National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

UNIVERSITY OF ALBERTA

WAFER LEVEL RELIABILITY FOR APPLICATION SPECIFIC INTEGRATED CIRCUITS

by

DWIGHT E. MANNING ©

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND
RESEARCH IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF ELECTRICAL ENGINEERING

FALL 1992

EDMONTON, ALBERTA

National Library of Canada
Canadian Theses Service
Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada
Service des thèses canadiennes

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-77102-X

Canadä

UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: **DWIGHT E. MANNING**

TITLE OF THESIS: **WAFER LEVEL RELIABILITY FOR APPLICATION SPECIFIC INTEGRATED CIRCUITS**

DEGREE: **MASTER OF SCIENCE**

YEAR THIS DEGREE GRANTED: **1992**

Permission is hereby granted to the UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the authors's written permisson.

(Signed) Dwight E. Manning

PERMANENT ADDRESS:
6925 F Rodling Drive
SAN JOSE, CALIFORNIA 95138
U.S.A.

Date: Sept 30 1992

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled WAFER LEVEL RELIABILITY FOR APPLICATION SPECIFIC INTEGRATED CIRCUITS submitted by DWIGHT E. MANNING, in partial fulfilment of the requirements for the degree of Master of Science in Electrical Engineering.

Don Koval

Supervisor: Dr. D.O. Koval

Keith Stromsmoe

Dr. K.~~E~~A. Stromsmoe

J C Sprague

External Examiner: Dr. J.C. Sprague

Date: Sept. 29 1992

This thesis is dedicated to the memory of
Allan William Edward Manning
for his loving support on the first one that led to this next one.

This thesis is also dedicated to my loving wife
Alison
for her loving support on this next one
and to my mother
Marian
for her loving support on both.

WAFER LEVEL RELIABILITY FOR APPLICATION SPECIFIC INTEGRATED CIRCUITS

ABSTRACT

Semiconductor designs and processes have undergone rapid changes in complexity, performance and size in a response to increased economic pressures to produce parts that are cheaper and more reliable than ever before in this highly competitive industry. New semiconductor processes, testing methodologies and procedures are being developed to increase the amount of reliability assurance testing through Wafer Level Reliability Testing, a new and emerging field in this industry, the subject of this thesis. The fundamentals of semiconductor manufacturing processes (e.g., photolithography, ion implantation and diffusion, thin film deposition, etching, etc.) are presented as part of a knowledge base required to understand the possible failure mechanisms, test structures and burn in reliability necessary to understand the application of Wafer Level Reliability Testing to custom ASIC (i.e. Application Specific Integrated Circuits) semiconductor devices and its importance to traditional life time reliability testing methodologies. Wafer Level Reliability (i.e., WLR) Testing fundamentals involving the fabrication and integration of test structures and production parts on a single wafer and specific tests for these test structures are presented. These unique WLR structures presented in this thesis for ASIC manufacturing have not been published to date in the literature. Empirical test results of traditional life time reliability testing for designed and fabricated wafers containing test structures and ASIC parts are presented. The empirical results are analyzed in detail to determine if there is any correlation between existing process monitor test and the reliability of the product to develop WLR lifetime models. Based on the empirical test results, the benefits and limitations of Wafer Level Reliability in controlling manufacturing processes of ASIC semiconductor devices will be discussed in detail in this thesis.

ACKNOWLEDGEMENTS

The author would like to thank Tom Long for the assistance and support in the beginning of the thesis and to thank Emery Sugasawara for supporting the completion of this thesis.

The author would also like to thank his supervisor, Dr. D. O. Koval, for all his guidance,support and insight during the preparation of this thesis.

The author would also like to thank and acknowledge the assistance of Brian Root and Tim Turner, the developers of the electromigration SWEAT test for their insights and updates on their technique.

The author would also like to thank the support of LSI Logic Corporation of Canada and LSI Logic Corporation for the support and the data enabling this thesis to be completed.

The author would also like to thank Mike Stover, Corporate Reliability Manager and Vish Bhide, head of the Wafer Level Reliability committee for their support. Joesph Zhou for his assistance in the data collection. The author would also like to thank his fellow Wafer Level Reliability team members for their contributions to the project : Ravindra Alluri, Sudhir Chopra, Abid Husain, Laique Khan, Daniel Gitlin, Napoleon Domingo, Liang Lie, Uresh Patel, Charles McDonald, Arthur Kuo and Bill Stevenson.

The author would also like to recognize the assistance of: Ross Manning and Xerox Canada for reproductions, Vivian Wells and Alison Manning for proof reading, Dr. C.R. and Arline James for transportation,Wayne and Heather Campbell for accommodation, and Ken Orr for the suggestion of a glossary.

TABLE OF CONTENTS

CHAPTER PAGE

I. INTRODUCTION 1

1.1 Yield and Reliability 2

1.2 Wafer Level Reliability 5

1.3 ASIC Circuits 5

1.4 ASIC Wafer Level Reliability 5

1.5 Thesis Objectives 6

1.6 Scope of Thesis 7

II. VLSI BASIC MANUFACTURING PROCESSES 9

2.1 Introduction 9

2.2 Clean Room Technology 10

2.3 Basic Semiconductor Processing 12

2.3.1 Thin Film Deposition 12

2.3.2 Photolithography 14

2.3.3 Etching 18

2.3.4 Ion Implantation and Diffusion 23

2.3.5 Overall Processing 30

2.4 ASIC Logic Circuit Basic Operation 37

III. YIELD AND RELIABILITY FUNDAMENTALS 45

3.1 Introduction 45

3.2 Defects 47

3.2.1 Defect Densities 51

3.3 Yield Modeling 51

3.4 ASIC Logic Circuits Test Methodology 54

3.5 Wafer Level Reliability Testing 56

3.5.1 Electromigration and Stress Void Reliability Failures............. 59

3.5.2 Transistor and Dielectric Reliability Failures 61
3.6 Reliability Defect Detection and Testing 62
3.7 Summary 65

IV. ELECTROMIGRATION FUNDAMENTALS 68
4.1 Introduction 68
4.2 Electromigration Physics 69
4.2.1 Structure of Conductors 69
4.2.2 Ion Potential Bonding and Diffusion 70
4.2.3 Ion Electromigration 74
4.3 Electromigration MTTF 78
4.4 Electromigration Test Techniques 80
4.4.1 SWEAT Test 81
4.4.2 BEM Testing 88
4.4.3 Other Test Techniques 91
4.5 Electromigration in VLSI Manufacturing 91
4.6 Summary 93

V. TEST CHIP DESIGN 95
5.1 Introduction 95
5.2 ASIC Manufacturing Requirements 96
5.3 Electromigration Test Structures 98
5.3.1 SWEAT Structures 98
5.3.2 Topology Test Structure 102
5.3.3 Contact and Via Test Structures 105
5.4 Dielectric Breakdown Test Structures 112
5.5 Hot Carrier Test Structures 118
5.6 Test Chip Layout 123
5.7 Summary 125

VI. BURN IN RELIABILITY 126

6.1 Introduction 126

6.2 Reliability Basics 127

6.3 Reliability Distributions 128

6.4 Failure Rate Measurements 133

6.5 Arrhenius Equation 134

6.6 Eyring Equation 136

6.7 Accelerated Reliability Testing 137

6.8 Semiconductor Reliability Testing 142

6.8.1 High Temperature Dynamic Lifetest 143

6.8.2 High Voltage Dynamic Lifetest 144

6.8.3 Low Temperature Dynamic Lifetest 144

6.8.4 High Temperature Storage 145

6.8.5 High Temperature / Humidity Lifetest 145

6.8.6 Temperature Cycling 145

6.8.7 Vibration / Shock Testing 146

6.8.8 Soft Error / Radiation Hardness Testing 146

6.9 Reliability Monitoring 146

6.10 Summary 151

VII. EVALUATION OF WAFER LEVEL RELIABILITY TEST DATA ..152

7.1 Introduction 152

7.2 Experimental Procedure 152

7.2.1 Test Run Processing 152

7.2.2 Test Wafer Layouts 157

7.3 Empirical Testing Results 159

7.3.1 Wafer Level Reliability Tests 160

7.3.2 Wafer Sort and Final Test 164

7.3.3 Reliability Stress Tests..165

7.4 Data Analysis 169
7.5 Wafer Level Reliability 174
  7.5.1 Statistical Process Control 176
  7.5.2 Product Lifetime Modeling 181
7.6 Summary 183

VIII. CONCLUSION **185**

IX. REFERENCES **189**

LIST OF TABLES

TABLE PAGE

3.1 Common yield and reliability defects 63

3.2 Wafer reliability tests 64

6.1 Acceleration factors for semiconductor reliability 141

6.2 Lifetime data 149

7.1 Wafer Level Reliability test runs 160

7.2 VRAMP normalized data 161

7.3 VRAMP normalized failure results 162

7.4 QBD normalized test results 163

7.5 BVOX normalized failure results 163

7.6 Normalized wafer sort yields 164

7.7 Normalized final test yields 165

7.8 Average normalized final test yields 165

7.9 Reliability stress test configuration 167

7.10 Reliability stress test results 168

7.11 Reliability stress test results summary 169

8.1 Semiconductor failure rates in different environments 188

LIST OF FIGURES

FIGURE		PAGE
1.1	Exploded view of a die on a wafer with test structures	2
2.1	Step 1 - metal / doped polysilicon deposition	14
2.2	Step 2 - photoresist deposition	16
2.3	Step 3 - photoresist exposure	17
2.4	Step 4 - photoresist develop	17
2.5	Step 4 - photoresist develop	20
2.6	Step 5 - metal/polysilicon etch	20
2.7	Step 6 - photoresist etch	20
2.8	First two steps in defining a hole	21
2.9	Steps 3 to 6 in defining a hole	22
2.10	First process step for low energy ion implantation	25
2.11	Process steps 2 through 5 for low energy ion implantation	26
2.12	Process step 6 for low energy ion implantation	27
2.13	Process steps 1 and 2 for high energy ion implantation	27
2.14	Etching process steps 3 to 6 for high energy ion implantation	28
2.15	Etching process steps 7 to 9 for high energy ion implantation	29
2.16	Cross section of a 3 layer metallized CMOS device	31
2.17	Basic CMOS cell - base array - one cell	32
2.18	Basic CMOS cell - contact holes defined	32
2.19	Basic CMOS cell - first metal layer routing	33
2.20	Basic CMOS cell - via holes defined	34
2.21	Basic CMOS cell - second metal layer routing	35
2.22	Basic CMOS cell - pad opening in the protective passivation	36
2.23	Cross section of CMOS transistors	37
2.24	N-Channel transistor with gate voltage at 0 volts	38
2.25	N-Channel transistor with gate voltage > 0 volts	39
2.26	Fully on N-Channel transistor	40

2.27 Basic invertor circuit 41

2.28 Double invertor circuit diagram 42

2.29 CMOS basic logic gates 43

3.1 Killing defect 48

3.2 Non-killing defect 49

3.3 Comparison of yield models 54

3.4 Potential reliability defect particle example 58

3.5 Stress void in a metal line 61

4.1 Crystal lattice structure 70

4.2 Ion self diffusion diagram 72

4.3 Types of Diffusion within a Thin Film 74

4.4 SWEAT structure resistance/temperature relationship 83

4.5 SWEAT power/temperature relationship 84

4.6 Typical SWEAT test structure 85

4.7 Histogram of BEM breakdown currents of production wafers 90

5.1 Gate array cell 96

5.2 Metal 2 SWEAT test structure 100

5.3 Metal 1 SWEAT test structure 101

5.4 Cross section of topology test structure with dielectric topology 102

5.5 Cross section of topology test structure with no dielectric topology 103

5.6 Topology electromigration test structure 104

5.7 N+ contact chain cross section 106

5.8 Metal spiking in N+ junctions 106

5.9 N+, P+ contact chain test structures 107

5.10 Minimum overlap N+, P+ contact chain test structures 108

5.11 Polysilicon contact chain test structures 109

5.12 Via chain cross section 110

5.13 Via test structures 111

5.14 Gate oxide test circuit 113

5.15 BPSG oxide test circuit **114**

5.16 Interdielectric test circuit **114**

5.17 Gate oxide and BPSG test capacitor structures **116**

5.18 Interdielectric test capacitor structure **117**

5.19 Electron flow in N-Channel MOS device **118**

5.20 Hot electron injection into N-Channel MOS gate oxide **119**

5.21 Hot carrier test circuit **121**

5.22 Hot carrier test structure **122**

5.23 Test chip layout **124**

6.1 Normal probability density function, f(t) **128**

6.2 "Bathtub" failure probability density function, f(t) **129**

6.3 Cumulative density function, F(t) of a normal f(t) function **130**

6.4 Reliability function, R(t) of a normal f(t) function......................... **131**

6.5 Component lifetime at two different reaction rates, R_1 & R_2 .... **138**

7.1 Test run foundry process flow **154**

7.2 Test run metallization process flow **155**

7.3 Test run assembly process flow **156**

7.4 Test wafer layout **157**

7.5 Example of Wafer Level Reliability scribe line test structures **159**

7.6 Wafer dielectric failure versus wafer sort yields **171**

7.7 Wafer failures versus gate oxide lifetime reliability failures **172**

7.8 Wafer failures versus final test yield **173**

7.9 SWEAT electromigration control chart.. **178**

7.10 Wafer Level Reliability Process Monitor flow chart **180**

GLOSSARY OF TERMS

Anisotropic:	An etch profile that is straight down or totally vertical with no horizontal etch.
Array:	A matrix of uncommitted logic functions contained in a single chip of silicon.
Assembly:	The process in semiconductor manufacturing where a completed die is put into a package or mounted directly onto a circuit board and connecting wire bonds are defined between the die and the package or circuit board.
ASIC:	Application Specific Integrated Circuit. This is a custom chip with an unique design. This part is not mass produced, and is generally referred to as a custom integrated circuit.
Base Array:	A gate array with the gates built up to the transistor level only. These wafers are not customized at this point and consist of standard sizes of chips with varying number of gates.
Bipolar Transistor:	The transistor type that works on current to switch or amplify.
BEM:	Electromigration test methodology that uses Median Energy to failure -MEF.

Burn In:	The acceleration of semiconductor parts to provide stress to shorten the lifetime of the parts such that information about the reliability can be obtained. Burn in usually involves heat as a stress and can be for infant mortality screening or product lifetime testing.
Chip:	A small piece of silicon that is a complete semiconductor device or integrated circuit.
Clean Room:	A room where the environment is strictly controlled and the level of airborne particulates is controlled through air filtration.
Contact:	A hole in the BPSG dielectric to allow first metal to connect to the transistor or doped silicon regions of a circuit.
CMOS:	Complementary Metal Oxide Silicon. Both N-channel and P-channel transistors are used together to form logic circuits. CMOS logic is low power.
Contamination:	Undesirable material. In semiconductor manufacturing, this refers to foreign matter in the circuit or chemicals/gases used in the manufacture of the circuits.
CVD:	Chemical Vapor Deposition - The process of depositing thin films on a wafer in a reactor by chemicals in gas form, reacting and growing a thin film on a wafer.
Defect:	A problem in a circuit on a wafer that can cause malfunction of the circuit.

Deposition:	The process in semiconductor manufacturing in which thin films are grown on a wafer.
DI Water:	De-Ionized Water is water that has been purified to the point where most of the ions usually present in water have been removed. This ultra-pure water is used throughout the manufacturing of semiconductor integrated circuits.
Die:	An integrated circuit that is in wafer form and is in the process of being manufactured.
Diffusion:	In semiconductor manufacturing, this refers to the process of using heat to allow movement of impurities through the semiconductor crystal structure and helps to define the conductive areas in transistors.
Doping:	The process of injecting impurities into the semiconductor crystal structure to help define the conductive properties of the device.
Downstream Etcher:	A type of plasma etcher, that has the gas plasma formed away from the wafer, so that the wafer is subjected to the minimum plasma damage. The wafer sits "downstream", closer to the vacuum pump, rather than closer to where the plasma is ignited.
DUT:	Device Under Test. This refers to the integrated circuit that is being tested by a computerized testing system.
ESD:	Electrostatic Discharge. Static damage is one way an integrated circuit can be damaged to cause failure in the circut.

Fab:	An industry common term that is an abbreviation for the Wafer Fabrication Facility.
FET:	Field Effect Transistor. This type of transistor uses voltage to build up a charge or electric field. This voltage is then amplified or used to switch on and off the transistor.
Final Test:	Upon completion of the assembly processes, the finished semiconductor product is subjected to a full functional test, fully testing the circuit to operational logic and functionality.
Foundry:	In ASIC gate array manufacturing, this refers to the initial processes in wafer manufacturing in which the silicon wafer has the transistor and logic gate structures defined up to the protective BPSG layer.
Gate Array:	A type of ASIC circuit that consists of logic gates.
HCMOS:	High Speed CMOS circuits.
Impurity:	In semiconductor manufacturing, this refers to atom(s), ion(s) or chemicals in the crystal structure of the semiconductor or a material that is not the material in question. Impurities in semiconductors are sometimes desirable. Impurities are used to define the conduction characteristics in the semiconductors.
Integrated Circuit (IC):	A circuit that is realized in a single semiconductor part.

Invertor Gate:	A basic logic gate that inverts the input or changes the input from high to low or from low to high.
Ion Implantation:	A semiconductor process that is used in manufacturing to put impurities into semiconductors to define conductive regions. A specialized machine is used to ionize impurities and accelerate them into the target; the wafer.
Isotropic:	An etch profile that is rounded. The etch rate of the vertical etch rate is the same as the horizontal etch rate.
Killing Defect:	A defect on a die that prevents the circuit from working.
Laminar Flow:	In semiconductor manufacturing, this refers to the continuous flow of filtered air in a clean room. The filters usually cover the entire ceiling area in VLSI manufacturing. This causes the flow to be everywhere within the clean room.
Line Yield:	This yield is the amount of good wafers or parts that are left after a particular process step(s).
Mask / Reticle / Photomask:	The plates that contain the patterns which are imprinted onto a wafer during the photolithographic process.
Masterslice:	Another term for a base array. A masterslice consists of uncommitted, unconnected logic functions only finished to the transistor level.

Memory Chip:	A chip that contains circuitry that can store data in the form of binary digits for later retrieval.
MEF:	Median Energy to Failure. The median amount of energy used in BEM electromigration testing that causes a test structure to fail.
Metallization:	The finishing processes in wafer manufacturing in which the metal interconnections of the uncommitted circuitry is defined.
Micron:	An unit of measure that is one-millionth of a meter.
Misalignment:	One or more layers of a chip are not lined up properly with the other layers. This is a mistake or defect caused in the photolithographic process.
MOS:	Metal Oxide Silicon transistor. An abbreviation for MOSFET, a type of field effect transistor. The metal-oxide-silicon refers to the physical construction of the transistor. There is an insulating oxide underneath the gate conductor.
MOSFET:	Metal Oxide Silicon Field Effect Transistor. See MOS
MTTF:	Median Time to Failure. The point in time where 50% of a population has failed.
N-Channel Transistor:	A type of MOSFET with N-type semiconductor source and drain regions.
N-Type Silicon:	Silicon with impurities implanted in the crystal structure, usually a type IV element, that gives an extra electron in the orbital shell.

NAND Gate:	A basic logic gate.
Nitride	An abbreviation for Silicon Nitride thin film which is used in the manufacturing of semiconductors.
Non-Killing: Defect	A defect that has not stopped the circuit from working.
NOR Gate:	A basic logic gate.
Oxide:	An abbreviation for Silicon Dioxide thin film which is used in the the manufacturing of semiconductors.
P-Channel Transistor:	A type of MOSFET with P-type semiconductor source and drain regions.
P-Type Silicon:	Silicon with impurities implanted in the crystal structure, usually a type III element, that gives an extra space for an electron or a "hole" in the orbital shell.
Parametric Test:	A test of the basic transistor circuit parameters usually done on specific test transistors and other associated structures in the scribe lines.
Parametric Tester:	A type of tester that is used to test out basic transistor or circuit electrical parameters.
Particle:	A piece of contamination that could end up on a die.

PECVD:	Plasma Enhanced CVD. The energy of the plasma gas is used to provide the energy for the chemical reaction.
Photolithography:	The process in semiconductor manufacturing of transferring the design from masks to wafers and defining the design in photo sensitive resist.
Plasma Etching:	The process in semiconductor manufacturing of etching exposed areas of materials through the use of excited reactive gases in the plasma state.
Process Monitor:	An in process test where tests are performed on wafers during the manufacturing fabrication processes. This can be done on dedicated test wafers or production wafers.
Reliability	In semiconductor manufacturing, this refers to how long the circuit will continue to work after manufacturing.
RIE:	Reactive Ion Etcher. This type of plasma etcher uses ion bombardment to assist the plasma etching. Wafers sit on the powered electrode.
Semiconductor:	A solid crystalline material (e.g. silicon) whose electrical conductivity is intermediate between conductors and insulators.
Scribe line:	The area between the die on the wafer where test structures, photolithography alignment targets and alignment measurement structures are located. This is the area also reserved for the width of the cutting saw when the wafer is cut up into individual die during the assembly processes.

Stepper:	A camera system which is used in the photolithographic process. This camera system only exposes a single or a few die at a time, called a field. After exposure of a field, the system moves or "steps" to the next location where it aligns to the previous layer(s) and exposes the same field in the next location.
Substrate:	A Silicon base upon which the logic functions are fabricated. The substrate influences the electrical characteristics of the transistors, which make up the circuit and also isolates the transistors from one another.
SPC:	Statistical Process Control. The application of statistics to build quality into a manufacturing process and product. Processes are reacted to before they are in an out of control condition. This is an entire manufacturing philosophy.
SWEAT	Standard Wafer-level Electromigration Acceleration Test - A test methodology and types of structure used in accelerated electromigration lifetime testing.
Test Structure:	A specific sub-circuit that is designed to test for one particular failure mode or to provide specific data. These structures can either exist on a die or in the scribe line area.
Thin Film:	A very thin layer of material deposited in a substrate, usually in the order of microns thick.
Transistor:	A semiconductor device that acts primarily as an amplifier or a switch.

ULSI:	Ultra large Scale Integrated circuits. This refers to the fabrication of circuits containing a high number of devices. Usually refers to over 100,000 logic gate devices.
Via:	A hole in the interdielectric material to allow one layer of metal interconnect to connect with another layer of metal interconnect.
VLSI:	Very Large Scale Integrated Circuits. This refers to the fabrication of circuits containing a high number of devices. Usually refers to over 10,000 logic gate devices.
Wafer:	A thin disk of semiconductor material in which many chips are fabricated at one time.
Wafer Fabrication:	The processes in which the desired integrated circuit is manufactured and fabricated into a wafer.
Wafer Sort:	The functional test that is done at the wafer level to determine which die are operating within normal specified parameters.
Yield:	The number of good devices or parts at the end of a process.

CHAPTER 1

INTRODUCTION

The manufacturing of Semiconductor circuits is a complex, highly technical process. Todays industrial trends are to fabricate more complex circuits with smaller and smaller geometries and with more complex combination of materials and layers. These Integrated Circuits, (IC) are manufactured on wafers. A wafer contains many die (i.e., die are called ICs in the final stages of fabrication) which are the individual circuits being constructed as illustrated in Figure 1.1 with various test structures for Wafer Level Reliability. All the die on a wafer are usually duplicates of the same circuitry and design.

Not all of the die on a wafer will function upon the completion of the manufacturing process. The number of working, functional die on a wafer is referred to as the wafer yield. For a completed circuit to work, all of the individual circuits that make up the die must be fully functional. A defect is a physical imperfection on a die that can cause part of a circuit not to work. A killing defect is a defect that actually prevents the circuit in question from working. In a working die, there must be no killing defects on the die. Defects do occur on die but are located such that they do not interfere with the operation of the circuit in question. These defects are called non-killing defects and do not reduce the yield of the wafer.

The wafer sort yield of a wafer is determined by fully functional testing of every die on the wafer with a tester and a test program upon completion of the manufacturing processes. After wafer sort testing, the ICs are tested again after each die has been packaged and the results are called Final Test. An IC is working, but for how long? Some non-killing defects can cause infant mortality reliability problems or early wear our reliability problems that significantly reduce the life expectancy of the IC. These failures are a major problem in the semiconductor industry

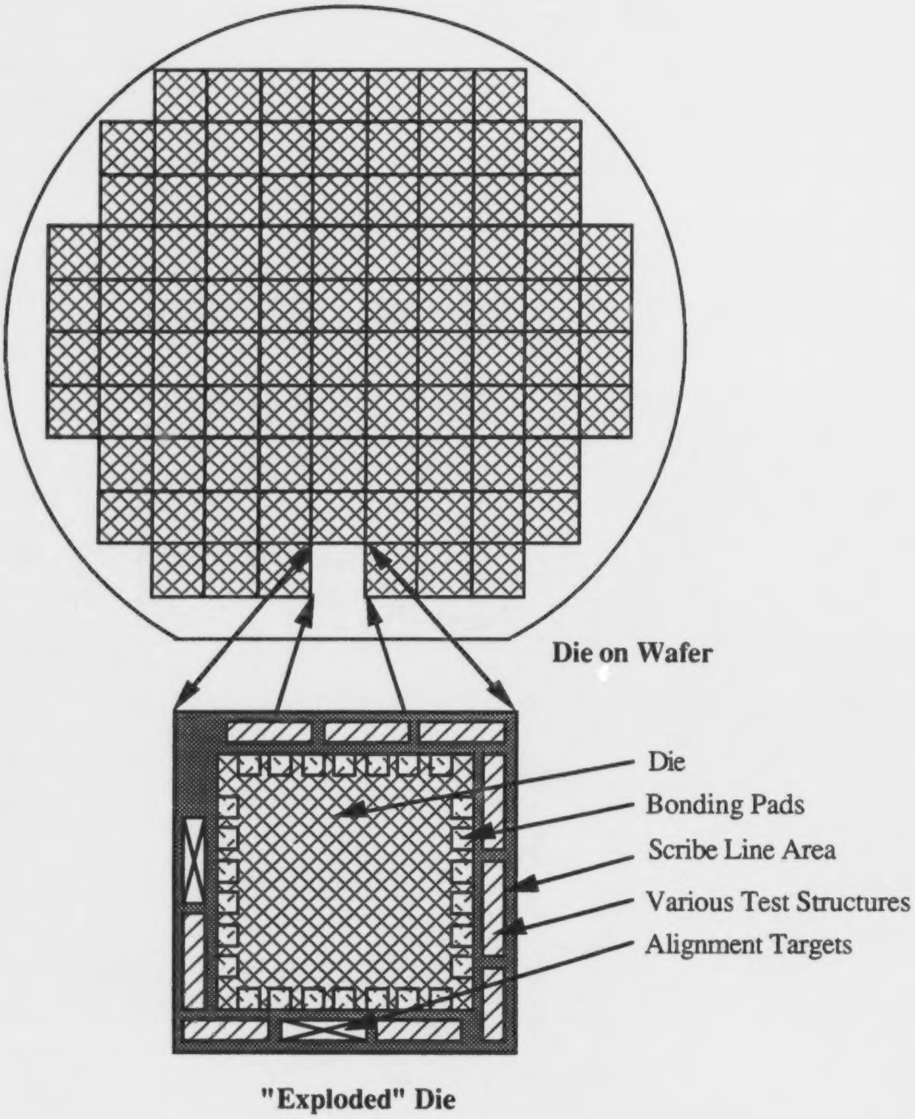

Die on Wafer
Die
Bonding Pads
Scribe Line Area
Various Test Structures
Alignment Targets
"Exploded" Die

Figure 1.1 Exploded view of a die on a wafer with test structures

1.1 Yield and Reliability

In semiconductor manufacturing, the yield and reliability of the parts have been traditionally treated as separate issues. It was believed that the defects that caused yield problems could be detected at the wafer sort step of the fabrication process and the reliability of the parts (i.e., individual ICs) was tested through burn in procedures. These procedures control infant mortality and collect data on circuit lifetime. The feedback from the burn in procedures was communicated to the engineers involved with the processing set up and control. The processing engineers tended to be more concerned with the immediate yield concerns and keeping their processes in control. Traditionally, reliability tasks involved initial burn in of finished parts to detect and weed out infant mortality failures and involved life time testing though accelerated stressing of the ICs and re-testing the ICs after the stressing. The wafers tended to be sampled from lots and these life time tests results were not known for a period of months. This delay prevented immediate feedback to correct out-of-control manufacturing processes.

But the industry has changed. Some of the assumptions that were valid five and ten years ago are now questionable. The majority of defects in the past tended to be infant mortality [1], e.g., similar to human life expectancy. The IC parts are now more sensitive due to the smaller geometries. The defect densities have decreased, enabling semiconductor factories to get yields on devices that were considered impossible several years ago. This has meant that there has been a shift towards more early wear out reliability failures. These early wear out failures (i.e., life expectancy greater than infant mortality) are becoming more dominant than the infant mortality failures that were seen in the past [1]. Because the dimensions are smaller, the parts are more susceptible to failures associated with current densities, materials, material stresses, etc.

Another issue of significance is that more processing is now done in a single wafer than ever before. The sampling plans for reliability were generally based on a lot by lot bases. Now the processes that individual wafers may see within a lot may have more variability than the variability in processing wafers from lot to lot. Some sampling plans that are based on lot variation may now be invalid. [1].

One of the biggest factors influencing the reliability picture of semiconductor devices is that customers are demanding higher quality than ever along with higher assurances of quality at an ever decreasing cost. To insure higher levels of quality through testing and yet continue to reduce the cost of the final product is an almost impossible task. To sample more ICs the traditional way that reliability testing was done costs more.

The answer? The semiconductor industry is working on incorporating more Wafer Level Reliability testing into the manufacturing processes. What Wafer Level Reliability testing refers to is the design of test and test structures that are sensitive to known physical attributes, that are carefully tested to control the stress variables and that promote a single primary failure mechanism while minimizing other failure mechanisms. These are placed on wafers that contain production die. These test structures are tested on a parametric tester. To test on a parametric tester, the structures and the tests have to be designed so that the tests can be performed quickly, usually in under a minute. The principle behind wafer reliability testing is the continuation of building in quality. It is being proactive rather than being reactive.

Traditionally, testing for product lifetime meant a life test through the use of five common types of stressing on the parts – temperature, voltage, current, humidity and temperature cycling [2]. Most product lifetimes are determined through temperature acceleration, based on the Arrhenius Equation [3]. Parts (i.e., ICs) are "burned in" to determine their lifetimes. Burn in refers to placing packaged parts in a burn in oven operated at elevated temperatures (e.g., 125℃), powering the parts up, and periodically testing the parts during the time terminated failure tests (e.g., common test procedures are terminated at the end of 1000 hours).

Upon failure, analysis is done to determine the defects that caused the failure. The procedure of burn in and analysis usually takes a period of months to complete. Most semiconductor manufacturers today adhere to an industry accepted military Standard MIL–STD–883 [4]. This involves a 1000 hour operating life test at 125℃. The problem is that the time involved here is a period of months. With the complex processing that is going on in todays semiconductor manufacturers, many unreliable products can be produced during this period before manufacturing problems are detected and rectified. This is a costly proposition for the semiconductor industry. The feedback from a single problem could take months. With Wafer Level Reliability testing, rapid feedback is now possible. Wafer Level Reliability will not replace the traditional burn in testing that is done for both infant mortality and long term life time testing, rather it will supplement this testing. It will provide additional data over a larger sample size that will ensure outgoing quality levels and provide quick feedback to control the on-going manufacturing process.

1.2 Wafer Level Reliability

Wafer Level Reliability is the testing of structures at the wafer level to determine reliability. Traditional semiconductor reliability testing has identified certain reliability failure mechanisms. Test routines and test structures have been proposed to test for these failures at a wafer level. These test routines and structures have been designed to allow for quick testing by a parametric tester. A parametric tester is a tester designed to test the electrical characteristics of semiconductor parts.

Wafer Level Reliability uses the combination of test structures and electrical stress to test for known reliability failures. The types of failures that these structures detect are: failure in metallization due to electromigration, stress voiding, contamination; failure in dielectrics due to trapped charges, radiation damage during processing, impurities, film stress; and failure in transistors due to hot electrons, ion impurities and improper processing.

1.3 ASIC Circuits

ASIC manufacturing is the manufacturing of Application Specific Integrated Circuits. Each product manufactured is a different chip. These are custom chips that are manufactured specifically for individual customers. This differs from other semiconductor manufacturers who manufacture standard devices where large quantities of a few designs are manufactured for multiple customers. There are basically two types of custom chips manufactured: array based and cell based [6]. Array based designs have the customized layers as the metal interconnection of circuits only. Array based designs have base array wafers. These base array wafers are used by multiple customers. They are only finished up to the transistor level and have the building blocks such as transistors, logic gates, memory, etc. are predefined. The designer customizes the circuit by specifying how these blocks are interconnected. Cell based designs are fully customized with the designer choosing which building blocks to use and the interconnection of them.

1.4 ASIC Wafer Level Reliability

Variation of customers designs results in variations in yield and reliability. A specific design on a array ASIC can have its own yield and reliability problems that are not seen on another design on the same base array. Todays ASICs are more complicated than they were ten years ago. ASICs now can contain on a single chip microprocessors, RAM, ROM, bipolar transistors, high current drivers in addition to the standard logic gates.

The problems with traditional Reliability testing through life time and infant mortality stressing are even greater with ASIC parts. There can be as few as two wafers run in a year for a particular design. ASIC manufacturers tend to have a great number of designs with just a few wafers each being manufactured for those designs at any given time, than do manufacturers of standard parts.

The traditional sampling plans for testing reliability failures through burn in are not adequate to account for all the different designs and circuits run in an ASIC plant. They tend to test for generic process failures. In the beginning of the ASIC manufacturing this was probably a valid assumption, but with the complexity of todays ASICs, there is a variation of physical structure and function between designs such that this may no longer be valid. The reliability of one design may not mean that another design is reliable.

The answer to this problem is to do testing on all designs. This can be economically accomplished by Wafer Level Reliability testing. However, there are still unique difficulties and challenges to doing Wafer Level Reliability on ASICs. Specifically, Wafer Level Reliability has been done on standard products in production lines that have few designs. ASIC production has hundreds of different designs. ASIC production involves the use of base arrays. The challenge is in adapting the Wafer Level Reliability test structures to utilize the existing base array structures.

1.5 Thesis Objectives

This thesis is directed at investigating Wafer Level Reliability, a key new area in today's semiconductor industry. To fully investigate Wafer Level Reliability on ASICs, the following thesis objectives will be presented and discussed in detail:

1. To present the basic VLSI/ASIC manufacturing process knowledge base necessary to understand the types and design of circuits being developed today, the types of failures encountered, the methods and equipment that test the parts and the types of reliability screening and testing used.

2. Adaptation of existing Wafer Level Reliability test structures to an ASIC Base Array.

3. Development of Test Routines to perform Wafer Level Reliability using a Parametric Tester.

4. Evaluate the results of Test Wafer Chips that were tested for Wafer Level Gate Oxide Reliability. These wafers contained real ASICs that will be subjected to traditional reliability burn in and lifetime tests.

5. Compare test structure lifetime reliability data with the production ASIC lifetime reliability data.

The Wafer Level Reliability test structures are not the same as the structure of the production IC circuitry. Answers to the following questions will be presented and discussed in some detail in this thesis: (1) Will the test structures accurately predict product life time? (2) Do the results of traditional reliability testing correlate with Wafer Level Reliability predictions?

Wafer Level Reliability structures and test methodologies will be adapted to the ASIC manufacturing environment. Structures, test programs and test methodologies will be designed to work within the constraints of ASIC manufacturing. The failures predicted by Wafer Level Reliability will be compared with the failures on ASIC product that have been subjected to life time burn in reliability testing. This thesis will investigate Wafer Level Reliability and evaluate its adequacy in predicting lifetime failures and provide immediate reliability failure mode analysis to control wafer fabrication processes.

1.6 Scope of Thesis

Chapter II will describe the basic VLSI manufacturing processes necessary to understand the nature of defects that cause failures in semiconductor circuits. Chapter III will present the fundamentals on yield and reliability concepts and definitions. Chapter IV will present the fundamentals of electromigration, a significant failure mechanism in very large scale integrated circuits. In this chapter a description of today's electromigration testing techniques will be presented. Chapter V presents the details of ASIC gate array Wafer Level reliability test chip design and layout incorporating electromigration, dielectric breakdown and hot carrier test structures. Chapter VI will present the basics of semiconductor reliability accelerated testing and monitoring. The fundamentals of the VLSI manufacturing processes, failure mechanisms, test chip design and burn in reliability presented in Chapter III to VI will provide a knowledge base for evaluating monitored processes and reliability test results. Actual normalized process monitoring and product reliability test results conducted on selected lots by LSI Logic Corporation will be presented and analyzed in detail in Chapter VII. Chapter VIII will present the conclusions and discussions of the thesis.

CHAPTER II
VLSI BASIC MANUFACTURING PROCESSES

2.1 Introduction

In modern manufacturing and fabrication of Integrated Circuits, the production of cost effective and reliable working parts is the key to success. Over the last twenty years, the fabrication of Integrated Circuits, (IC) has gone from a fledgling industry to one of the most important and demanding manufacturing technologies in the world today. Today's ICs are larger, more dense, faster and require a higher level of manufacturing technology than even the circuits manufactured five years ago. To keep competitive, a manufacturer must have good yields. This means that a manufacturer must always try to maximize the number of good reliable parts. The yield is nothing more than the ratio of good parts to the total parts manufactured. This is not as simple as it sounds. With the increasing demands on performance, quality, cost and reliability of the finished ICs, yield becomes extremely important and harder to maintain. The state-of-the art in terms of manufacturing continues to change quickly in the industry.

To properly examine what yield and reliability is in today's Integrated Circuit manufacturing technology, one must have a basic understanding of the manufacturing processes involved. There must be an understanding of the types and design of circuits being manufactured, the types of failures encountered, the methods and equipment that test the parts and the types of reliability screening and testing used. This interrelationship is becoming more complex. With the demand to make circuits better, faster, and with higher performance, understanding their interrelationships are important to understanding why integrated circuits fail. A basic overview of some of the processing steps is important to understand where circuit failures can occur.

2.2 Clean Room Technology

In today's Integrated Circuit (IC) manufacturing technology, ICs are manufactured in factories that contain clean rooms. A clean room is a room which has a controlled environment that is compatible with the requirements for manufacturing of Very Large Scale Integrated circuits (VLSI) and Ultra Large Scale Integrated circuits (ULSI). Most manufacturing of VLSI and ULSI involves dimensions in the one micron and sub-micron regions. Therefore, controlling the environment is critical in obtaining working circuits. To what level is the environment controlled? The level of control of the environment is partially dependent on the processing steps that are being performed in the particular clean room or clean room module in question. Most clean rooms are controlled by microprocessor controller systems that monitor and adjust the environment in the clean rooms to a high degree of accuracy.

Traditionally, clean rooms were rooms that had just the level of airborne contamination minimized and controlled. By controlling the level of airborne contamination, the number of defects due to these particles of contamination getting into the circuit is reduced. Airborne particulate levels are controlled by having a laminar air flow within the room. Motor blower units force air through Hepa filters. These Hepa filters filter particles of size 0.1μm and larger. By having a laminar flow of air, the rooms tend to clean themselves and any airborne particulates are swept away in the air flow. These particulates are eliminated.

But the cleanliness of air is not the only factor that must be controlled in order to make a clean room "clean". The materials that go into construction, the tables used, the process equipment and any other material must be clean room compatible. That is, all objects, including people, in a clean room must be clean room compatible. To be clean room compatible is to be particulate free. All tables must be constructed as to not hamper the flow of the clean air. All people in the clean room must wear garments that are not particulating called clean room garments or "bunny suits". Even though it is impossible to entirely eliminate contamination within the clean room, the goal is to minimize the sources of contamination as much as is practical.

In addition to cleanliness, there are other environmental factors that must be controlled in VLSI and ULSI fabrication facilities of today. Temperature is a factor in many processes. In most photolithography areas, the temperature must be controlled to ±0.5 °C to keep the sensitive photolithography camera systems in focus. Humidity is also an important factor. Photoresists used in today's process can absorb moisture or dry out changing the photo speed of the resist and also changing its ability to resist etching processes. As has been seen, a clean room is not only a room that has Hepa filtering and laminar flow, it has evolved with the technology to become a controlled environmental area that is conducive to the manufacture of large circuits with very small structure dimensions.

But controlling the environment inside the clean room is only part of the challenge. The facilities that supply the clean room and the process equipment must be of a very high purity level. These can be another source of contamination. Some examples of what these facilities are: Clean dry air, Pure Nitrogen, Process Gases, De-Ionized (DI) Water and House Vacuum. Clean dry air is used to run pneumatics on the various process equipment used in the manufacturing of ICs. Pure Nitrogen is usually 99.9999% pure and is used in almost all processing equipment for venting from vacuum, purging, etc. DI Water is De-Ionized Water and has been filtered and purified through reverse osmosis and passed through resin beds. The water's purity is measured by its resistivity which for VLSI and ULSI is 18 MΩ (Megaohms) pure. The house vacuum is a source of vacuum to hold wafers in place during regular process steps.

In addition to these facilities, there are chemicals and ultra pure gases that are used in the processes themselves. These have to meet stringent requirements as to their composition and as to their level of contamination or impurities. The process equipment itself must be designed and constructed to very strict requirements. They must perform the processes in question, be compatible with the clean room environment and handle the wafers in such a fashion as to minimize the amount of contamination on the wafers.

The demanding requirements of VLSI and ULSI manufacturing technology puts high demands on the environment and processes that are used in this manufacturing. The clean room and its facilities are as important as the processes themselves. If an essentially contamination free environment and contamination free processes do not exist, then there will be very few working parts out of the factory. Clean rooms today have to provide this very high level of environmental purity and control to be able to manufacture the highly complex circuits.

2.3 Basic Semiconductor Processing

In order to properly understand the nature of defects that cause failures in semiconductor circuits, there has to be a fundamental understanding of the basic processes involved in the manufacture of these circuits. As the complexity of the circuits increase, so do the amount of process steps required to manufacture the circuits. With more steps, there are more chances for problems, mistakes and misprocessing to occur. To understand the failures caused by these difficulties, one must understand how the basic process steps are performed.

In Integrated Circuit (IC) manufacturing, circuits are constructed in a layer by layer process. For example, a metal interconnect layer is first constructed before an insulator layer with via holes is constructed. One layer is finished before another is built. Even though there are many steps to building a completed wafer with working circuits, most of the steps fall into four basic processes: Thin Film Deposition, Photolithography, Etching, and Ion Implantation and Diffusion. A basic understanding of the fundamentals of these processes is important to understanding the failures that can occur because of these processes.

2.3.1 Thin Film Deposition

Thin film deposition is one of the "four" basic processes. The thickness of films used in semiconductor technology usually ranges in thickness from hundreds of Angstroms to a few microns. Hence the name, "thin films" because of the small dimensions. Thin films have unique properties because of their very thin nature. This allows atoms and ions to move around and diffuse easily. Properties such as this in thin films can cause yield and reliability problems. Many factors have to be controlled in the thin films that are deposited. Factors such as film composition, film stress, film grain size, thickness of the film, purity of the film, uniformity of the thickness, and the film density can all be important. The critical factors are dependent upon the process, circuit design and the intended use of the film in question.

Wafers are constructed in a layer by layer process. The layers are built up through subsequent depositions. The layers have different materials deposited to give the desired properties that each individual layer requires. For example, polysilicon and metals are used for conduction, oxides and nitrides for insulation.

Oxidation is one method used to grow thin films, especially in Silicon MOS technology. Silicon dioxide, SiO_2, is grown thermally in the presence of oxygen on silicon wafers. Silicon dioxide is commonly referred to as "oxide" and silicon nitride is commonly referred to as "nitride" in the industry. Three things are needed for oxidation of a silicon wafer - a source of oxygen, energy (usually thermal energy in the form of heat) and a silicon substrate. Oxygen comes from using oxygen gas or water vapour. With the assistance of heat, the oxygen diffuses through the already grown oxide and reacts with the silicon to form more oxide. Silicon is consumed in the process.

Chemical Vapour Deposition is another method of depositing thin films. Films are grown in a reactor by having process gases react and form on the wafer. How the reaction takes place depends on the type of reactor and the film grown. There are three basic types of reactors used: atmospheric chemical vapour deposition (CVD), low pressure chemical vapour deposition (LPCVD), and plasma assisted chemical vapour deposition (PECVD). LPCVD is done under a vacuum, hence the name "low pressure". PECVD uses radio frequency power to ignite the process gases into a plasma state. Heat was usually used to provide energy for the reactions to take place. In PECVD, the plasma provides the energy for the reactions. The reactor used depends upon the film grown and the process goals.

Evaporation is a traditional technique used in the deposition of metals on wafers in semiconductor processing. Metals that have been commonly used are aluminium, copper, tungsten and gold. Evaporation is a technique that uses a high vacuum chamber. The metal to be deposited is heated in this high vacuum. The metal is heated through a number of different means and is referred to as the source. The metal evaporates in the high vacuum and is deposited on the wafers.

Another technique for deposition of metals is called sputtering. In this procedure, ions are accelerated through a high vacuum and strike a target. The target is made up of the metal that is to be deposited. The momentum of the ions are transferred to the atoms of the material and the material is transferred in vapour form to the wafers. The ions act as "bullets". Argon is typically used in sputter systems to act as the accelerating ions, or "bullets". The ions are accelerated through electric and or magnetic field attraction. The metal can be titanium, platinum, gold, tungsten, copper, aluminium, etc. This procedure is a purely physical phenomenon, involving no chemical reactions. To define a line in a material such as polysilicon or aluminium, the material is deposited on the previous layers or substrate first. The result of this deposition is shown in Figure 2.1.

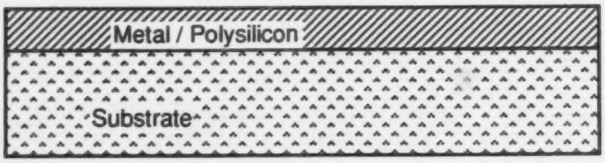


Figure 2.1 Step 1 - metal / doped polysilicon deposition

Through all of these techniques, the materials or thin films that are used in the manufacture of ICs are deposited on wafers. The uniformity and control of the composition of these films are critical. Any impurities or mistakes in either the composition or deposition of these films may cause defects that will affect the yield and reliability of the finished ICs.

2.3.2 Photolithography

Photolithography is the process used in IC manufacturing to transfer the design for a particular layer into photoresist on a semiconductor wafer. Since the majority of manufacturing done in the world today is on optical steppers, this is the type of system that will be described here. The principles are the same for projection aligners, direct write Electron Beam systems and other pattern transfer systems. The goal of photolithography is to accurately reproduce on the wafer the circuit layout in photoresist for the particular layer in question. The photolithography step is repeated many times throughout the process. The photoresist is used to resist etching materials or to prevent underlying layers from being damaged or doped during ion implantation.

The pattern that is to be reproduced is usually on a glass and chrome plate and this plate is referred to as a mask. A mask can consist of one layer of a single die or one layer of a few die. Some masks even have multiple layers on them, but all masks have only one layer used at a time. It depends on the size of the die being manufactured and the photolithographic camera equipment for the number of die exposed at one time. Most steppers today have a maximum field size of 1.5 cm by 1.5 cm. Some new steppers can go as big as 2.0 cm by 2.0 cm. A field is a region on a mask that consists of a single layer for one or more die that is exposed in a single exposure by the camera system. The size of the fields are limited by the lens technology.

A stepper is a camera system that operates in the following fashion. The stepper first finds alignment targets that have been defined at a previous layer. It then uses these targets to line up the layer with the mask that is to be exposed. When the alignment is within the machine's alignment tolerance, an ultraviolet light is projected through the mask and onto the photoresist on the wafer, thus exposing the wafer. The stepper system then "steps" to the next alignment site. One or a few die are exposed at each exposure.

The reason why the best steppers in the world today only have a field size of 2cm by 2cm is because this is the area in which the stepper manufacturers guarantee minimum distortion. Distortion in the lens will cause errors in the reproduction of the mask image on the wafer. This distortion can cause various defects in the finished IC.

Even though an individual lens element within a lens system can not be manufactured without distortion, a particular lens element can be manufactured with a known distortion and another lens element with the opposite lens distortion. These two particular lens elements may be part of a lens system that makes up the stepper lens. The cumulative effect of these two lens in a multi-lens element system is to minimize the distortion in question. It is through this highly specialized multi-lens element lens system that the effects of optical distortion are minimized. In todays leading edge stepper technology, the lens can consist of up to twenty elements. The design of these precision optics is another very sophisticated high technology field. The larger the field, the harder it is to manufacture a lens system free of defects.

It is important to understand some of the basics of stepper technology in order to understand some of the defects that are caused by mistakes in the photolithographic processes. If there are any problems on the mask or with the operation of the stepper, this can affect every die on all the wafers being run. The defect is not limited to a single or few die but is repeated through the stepping and exposing action of the stepper camera system. This kind of defect is referred to as a "repeating defect".

The photolithographic process consists of three basic steps - photoresist deposition on the wafer, exposing of the desired pattern on the wafer and developing and hardening the patterned photoresist that is on the wafer. A layer of photoresist is deposited on the wafer in a very uniform fashion. This photoresist deposition is usually done in the following fashion. The photoresist or resist is delivered in a solvent form to a spinning system. The wafer sits on a spin chuck and photoresist is deposited on the wafer. Through the action of spinning, the excess resist is spun off and a very uniform controlled thickness of photoresist is left on the wafer. The wafer is then baked to drive off the excess solvents. With too many solvents, the under exposed areas may mix with the exposed areas after exposure. By using the cross section diagram shown in Figure 2.1, the photolithography process of transferring a line pattern from a mask to photoresist can be illustrated. If we take the same substrate that was used in Figure 2.1 with its thin film on it and spin the photo resist on top, the cross section now will look like Figure 2.2.

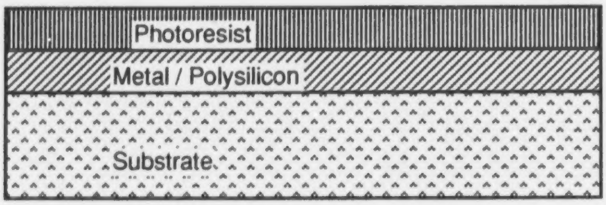


Figure 2.2 Step 2 - photoresist deposition

Masks for the layer in question are loaded into the stepper system. The wafers are exposed as described above. Photoresist (i.e., called resist) undergoes a molecular transformation called polymerization when exposed to ultra violet light. Exposed areas under go a chemical change. During developing, only the exposed pattern remains. If we return to our example of defining lines in a thin film material, the process of exposing the photoresist is illustrated in Figure 2.3.

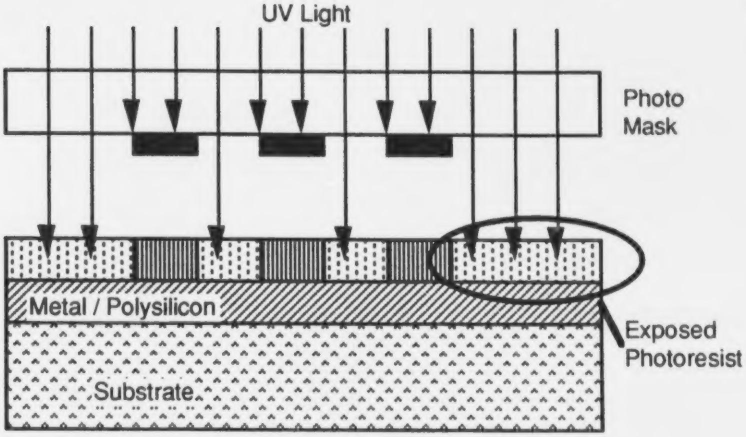

UV Light
Photo Mask
Metal / Polysilicon
Exposed Photoresist
Substrate

Figure 2.3 Step 3 - photoresist exposure

Through the process of developing, the undesired areas are washed away leaving only the desired pattern on the wafer in photoresist. The wafers then go through a hard bake process that hardens the photoresist. Hardening the photoresist makes the photoresist more resistant to the etching and implantation processes. Again after the developing, the cross section of our example now has the lines defined in the photoresist. This is illustrated in Figure 5.

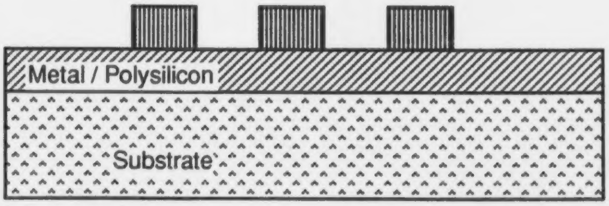

Metal / Polysilicon
Substrate

Figure 2.4 Step 4 - photoresist develop

The processes described above is a basic description of the photolithographic processes. There are additional types of pattern transfer techniques that wafer fabrication factories may use, depending upon the application. The overview given above still applies in these more "specialized" cases. Most photolithographic processes in production today are based on this "basic description."

2.3.3 Etching

Etching is the method of removing materials in semiconductor processing. Etching usually transfers the pattern that has been defined in photoresist into the material in question. The etching processes that do not transfer the patterns are usually blanket removal etches, such as the blanket removal etch of the photoresist.

Today, most etches are done with the use of excited gases through plasma etching. Plasma etching refers to the state of the etching gases. In a vacuum, etchant gases are excited through the application of electrical RF energy. The gases are in an excited, highly energetic state. The gases tend to exist in this plasma as single ions and atoms. These chemical species are highly volatile and react with the material placed in the etcher, such as the semiconductor wafer.

Different etchant gases are used for different plasma etches. For example, chlorine is used for aluminium and polysilicon, freons for oxide and nitride, and oxygen for photoresist. The goal of the plasma etching is to have the etchant gas in a plasma state, have the reactive species react with the desired material, and have the reaction by-products be gaseous so that they can be easily removed from the etching chamber. The material to be etched determines what chemical reactions will take place and therefore, what etchant gases will be used to form the reactions in the plasma etch chamber.

Plasma etching is not only the chemical reactions. With the application of RF electrical energy, ions are accelerated towards the electrodes in the system. The electrical bias on the electrodes will determine the amount of acceleration. Basically there are three types of plasma assisted etchers. The different types of etchers are defined by where the wafer is located in the etching system. Reactive Ion Etchers (RIE) have the wafer sit on the powered electrode. Here, the biases that are seen are usually in the hundreds of volts region. This high bombardment is generally used to etch species that are difficult to etch, such as aluminium. Aluminium is very difficult to etch because a native oxide, aluminium oxide, is formed on contact with oxygen in the air. This layer is very unreactive and requires high bombardment with the ions to be removed.

Plasma etchers are plasma assisted etchers that have the wafer on the grounded electrode. Here tens of volts of bias are used to accelerate the ions. These are lower ion bombardment machines.

The etch rate is more dependent upon the chemical reaction in a plasma etcher than in the RIE systems. They tend to run at higher pressures. The higher pressure means more reactants. This is how the plasma etchers etch materials as fast as the RIE systems.

Another type of etchers used in manufacturing of semiconductors is the down stream etcher. The wafers are "down stream" from where the plasma is generated. The excited etch species are created between the two electrodes. These species are then transported to the wafer where the reaction takes place. The wafers are situated near the vacuum source, and hence "down stream" from the area where the plasma gas is generated. The goal here is to have no ion bombardment and the reactions that take place are purely chemical in nature, and not ion assisted.

Before down stream etchers, purely chemical etching was done with the use of wet chemicals. Wet chemical etching involves baths in which wafers are submerged in the etchant solution or machinery that "sprays" the etchant solution onto the wafer. The wet chemical processes are generally harder to control and dirtier, but this does not mean that they are useless in processing. They do not cause any damage to the wafers through ion bombardment. Wet chemical etching is cheaper than plasma etching. This is how wafers were traditionally processed before plasma etchers were available. Wet chemical etching is still in use in most manufacturing processes today, usually in non-critical applications or applications that are sensitive to damage that a plasma assisted etcher may cause.

Most etching processes that define structures involve at least two etch steps. The first step involves the etching of the desired material. The pattern is transferred into the material because the photoresist protects the areas that are not to be etched. The second step of this etching process is the etching of the photoresist layer. Once the photoresist layer is removed, the only place where underlying material is left is where the photoresist was. This structure corresponds to the circuit design for that layer and the pattern that was on the mask. For example, metal lines are defined in aluminium by first etching the unprotected aluminium and then etching the photoresist that protected the aluminium. After this step, there are the metal conductor lines left that correspond exactly to the lines on the mask and in the design. This is how the pattern is transferred from the exposed photoresist to the material layer. To illustrate this process, the example of lines defined in the photoresist defined in Figure 2.4 is repeated in Figure 2.5

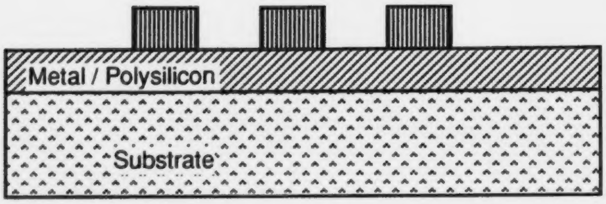


Figure 2.5 Step 4 - photoresist develop

After etching, the cross section is altered to that shown in Figure 2.6

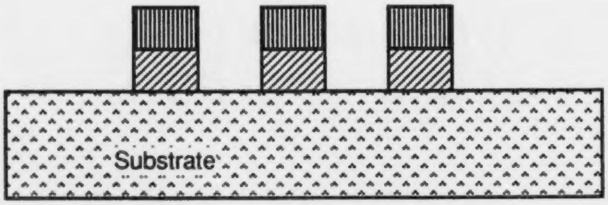


Figure 2.6 Step 5 - metal/polysilicon etch

The lines are now defined in the thin film (metal/polysilicon) to complete the process. The photoresist is then etched off or stripped. The final cross section is shown in Figure 2.7 below.

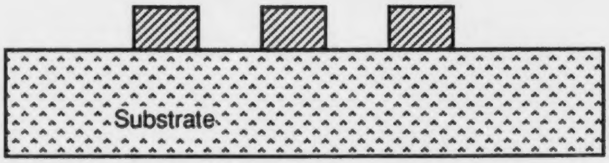
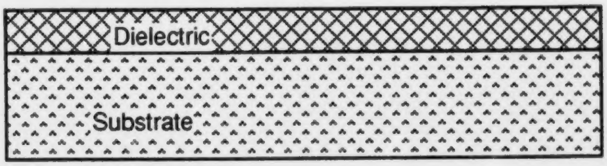


Figure 2.7 Step 6 - photoresist etch

The goals of the process determine which type of etcher to be used. Plasma assisted etchers allow the more precise control of etch rate, uniformity of etching across the wafer, the etch profile and the selectivity of etch rates. The selectivity refers to the ratio of etch rates between two different materials. For example, the selectivity of aluminium etch rate to oxide etch rate is the ratio of these two etch rates. This is important in the etching of aluminium lines. Aluminium or its alloys are a common conductor used to interconnect circuits together in semiconductor processing. This is usually deposited on an insulator such as silicon dioxide, or "oxide". The goal of etching metal lines is to etch the metal and not the underlying oxide. Therefore, the selectivity of aluminium etch rate to oxide etch rate should be as high as possible, to minimize the etching of the underlying oxide. Etching too much oxide can cause the aluminium lines to lift or provide topology related problems at subsequent layers.

In the previous example, line were defined in a thin film such as a metal or polysilicon. The process is similar for holes in a dielectric such as silicon dioxide or silicon nitride. The entire process of defining holes such as contact or via holes is illustrated in the sequence of steps shown in the diagrams of Figures 2.8 and 2.9, respectively..

Step 1 - Dielectric Deposition



Step 2 - Photoresist Deposition

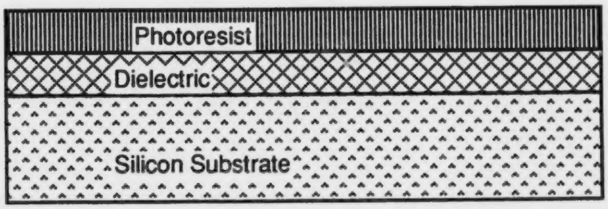
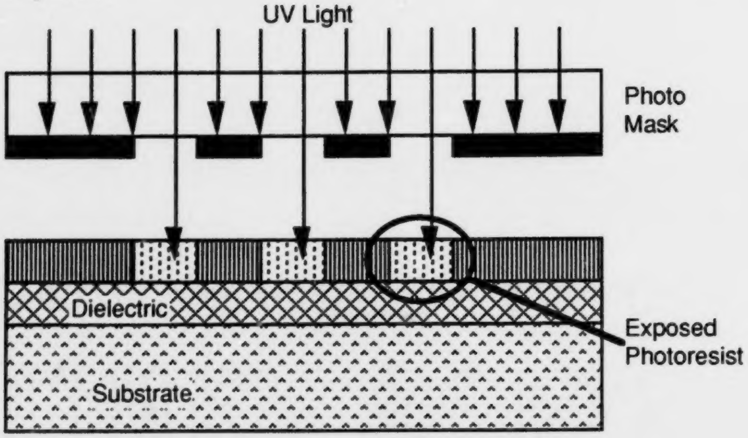
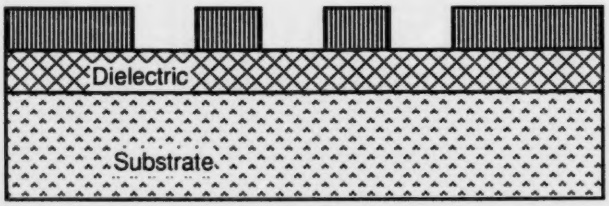


Figure 2.8 First two steps in defining a hole

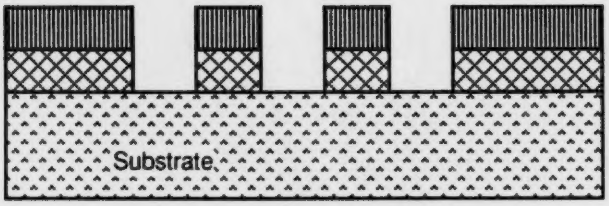
Step 3 - Photoresist Exposure


UV Light
Photo Mask
Dielectric
Substrate
Exposed Photoresist

Step 4 - Photoresist Develop


Dielectric
Substrate

Step 5 - Dielectric Etch



Step 6 - Photoresist Etch

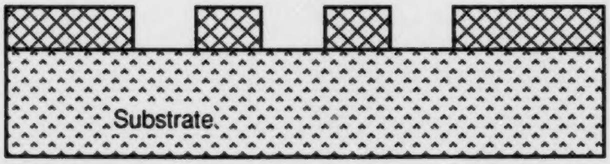


Figure 2.9 Steps 3 to 6 in defining a hole

The etch profile refers to the cross section profile of the etched material. The profile is determined by the etch rate vertically (depth) as compared to the etch rate horizontally (sideways). If the etch rate is the same for both the vertical and horizontal, the etch is isotropic. If the vertical etch rate is essentially all vertical the etch is said to be anisotropic. Wet chemical etching and down stream etching gives isotropic etch profiles. RIE and Plasma etchers tend to give anisotropic etch profiles. Although the down stream etcher does not have the ion bombardment and therefore does not damage underlying structures, it can not give the desired profile control or etching characteristics desired in certain types of etches. Therefore, some damage due to ion bombardment is traded for the control of the profile and the control of the etching. The damage caused may not be critical to either the yield or the reliability of the circuit, depending on the etching processes in question.

2.3.4 Ion Implantation and Diffusion

Ion Implantation and Diffusion are used in semiconductor manufacturing to define impurities in silicon to make n and p type silicon. The types of impurities, the amount of impurities, and the depth of implantation all give rise to the desired conduction properties that make up transistors, memory, CMOS chips, etc. In terms of understanding the basic processing, what is being manufactured is not important as to understanding the mechanisms behind Ion Implantation and diffusion. All types of semiconductor manufacturing require some form of Ion Implantation and Diffusion. Impurities that define the characteristics of n or p type semiconductors are formed by either diffusion or ion implantation with diffusion.

Semiconductors are of a type IV elements (i.e. Si or Ge) in the periodic table. Impurities are usually of the type V or III. Typical impurities used in forming n or p type silicon are arsenic, phosphorus, and boron. Impurities either give an extra electron to the crystal lattice structure (N type semiconductor = type IV Element) or have a hole where an electron can go (P type semiconductor = type III Element). The impurities and semiconductors used may not necessarily be elements. Compounds with the same electron configurations as the elements mentioned above can act as impurity dopants. Other compounds can act as semiconductors, depending upon the compound's electron configuration. An example of a semiconductor compound is Gallium Arsenic.

Diffusion in semiconductors is the most widely used process to control the depth of impurities into semiconductors. Impurity atoms move under the high temperature diffusion processes. Basically, the rate of diffusion is controlled by the amount of impurities, the temperature used, and the time that the wafer is exposed to the high heat. Impurities are introduced to the surface usually through contact with the impurity in vapour form or through ion implantation.

Ion implantation (i.e., a technique where ions are accelerated to a high speed and shot into the wafer surface) is used in semiconductor manufacturing to introduce impurity ions into semiconductors. The advantage to using ion implantation is that the concentration and profile of the implantation can be controlled precisely. This is why ion implantation is in use through out the industry today. To do ion implantation, wafers are put into a process machine called an ion implanter. The wafers are located in the target position. The impurity used are energized under vacuum to form ions. These impurity ions are then accelerated through use of a magnetic field and focused on the target (wafer). The ions have enough energy to act like bullets and penetrate into the material. Control of the acceleration energy gives the control of the depth that the ions penetrate. Because ions are charged, the current of the ion beam can be measured. By knowing the time the beam is on and the area that is implanted, the amount of implanted charges can be calculated. If the dose is referred to as f, then the dose is given by the following equation: [7]

$$\phi = \frac{Q}{mqA} \quad \text{atoms/cm}^2 \qquad \textbf{(2.1)}$$

where:
m=charge state of the ion (+1 or -1)
q=electron charge
A=Area
Q=integrated charge

and if the integrated charge Q is given by: [7]

$$Q = \int I \, dt \qquad \textbf{(2.2)}$$

where: I (i.e., the beam current) is applied for t seconds.

By knowing the area the beam exposes, the time of exposure and the beam current, the dose is easily calculated and accurately determined. Once the dose is known, it is then easy to accurately control the doping at any depth. This gives accurate control of the doping profile.

Since the ion implantation is done under vacuum conditions, it tends to be a much cleaner process than other impurity inducing processes. Most processes in semiconductor manufacturing involve ion implantation followed by a high temperature diffusion step. This allows accurate control of both the deposition and depth of implant. For deep implants, the diffusion step is necessary because ion implantation alone becomes harder, more difficult to control and more expensive, with the deeper the implant required. A shallower implant followed by a diffusion step is easier and more economical than an implant alone. Areas that do not receive ion implantation are protected through putting a protective material over the area. This area is defined through photolithography or "masking". Photoresist is used to protect specific circuit areas from low current implants. The areas that require masks are exposed and developed leaving holes in the photoresist for the implantation.

To illustrate the ion implantation, the sequence of steps necessary to process a low energy ion implantation of a shallow junction such as a voltage threshold adjust implant (VT) is shown in Figures 2.10, 2.11 and 2.12.

Step 1 - Photoresist Deposition

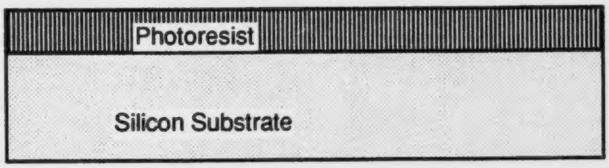
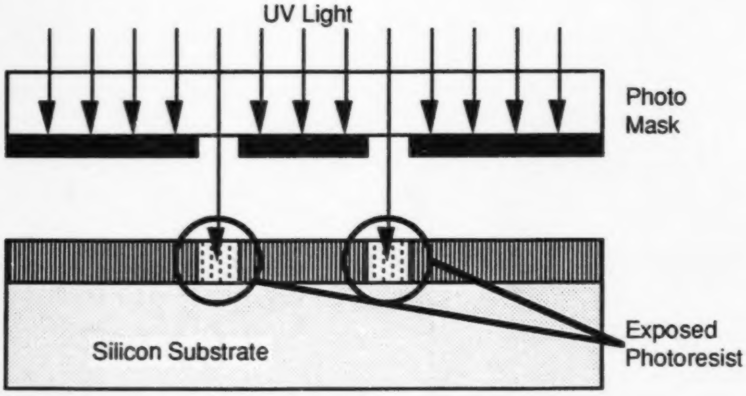
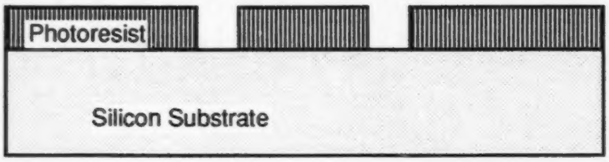


Figure 2.10 First process step for low energy ion implantation

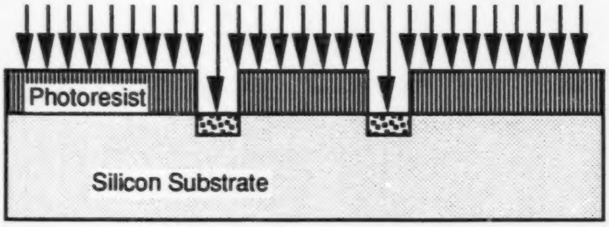
Step 2 - Photoresist Exposure


UV Light
Photo
Mask
Exposed
Photoresist
Silicon Substrate

Step 3 - Photoresist Develop


Photoresist
Silicon Substrate

Step 4 - Low Energy Implantation


Photoresist
Silicon Substrate

Step 5 - Photoresist Etch (Strip)

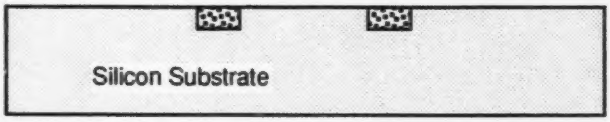

Silicon Substrate

Figure 2.11 Process steps 2 through 5 for low energy ion implantation

Step 6 - Junction Diffusion

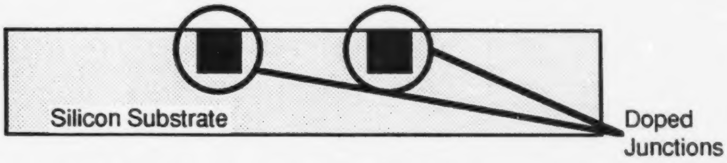
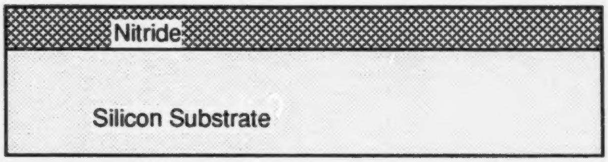

Silicon Substrate
Doped Junctions

Figure 2.12 Process step 6 for low energy ion implantation

For higher current implants, a deposition of a protective material such as silicon nitride is used. The nitride is then masked by photoresist and the exposed area is etched away. The nitride then forms a protective mask that protects certain areas from the implant. After implant, the nitride mask is etched away. This is illustrated in the example of a high energy implant as shown in the sequence of diagrams shown in Figures 2.13, 2.14 and 2.15.

Step 1 - Nitride Deposition


Nitride
Silicon Substrate

Step 2 - Photoresist Deposition

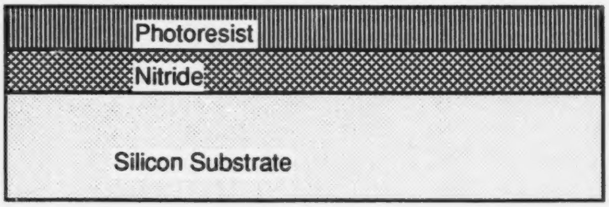
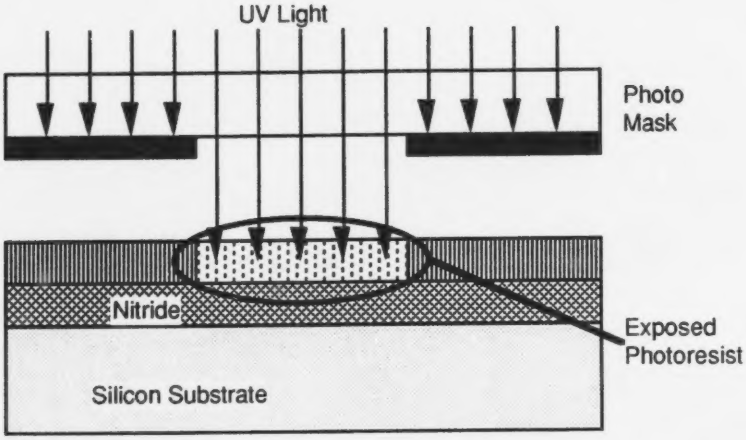
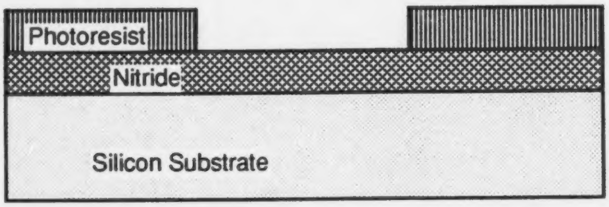

Photoresist
Nitride
Silicon Substrate

Figure 2.13 Process steps 1 and 2 for high energy ion implantation

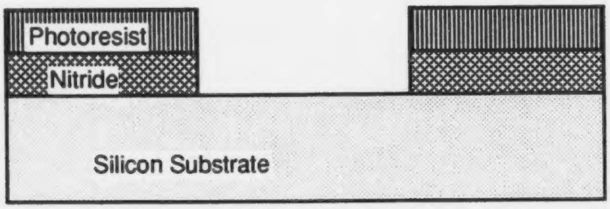
Step 3 - Photoresist Exposure


UV Light
Photo Mask
Nitride
Exposed Photoresist
Silicon Substrate

Step 4 - Photoresist Develop


Photoresist
Nitride
Silicon Substrate

Step 5 - Nitride Etch


Photoresist
Nitride
Silicon Substrate

Step 6 - Photoresist Etch

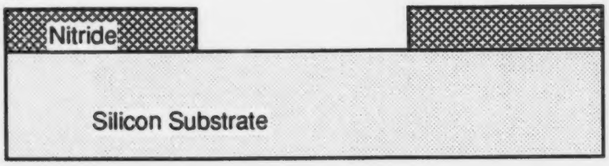
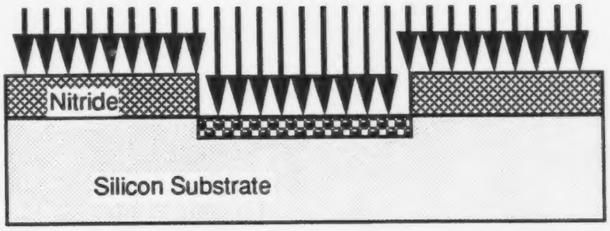
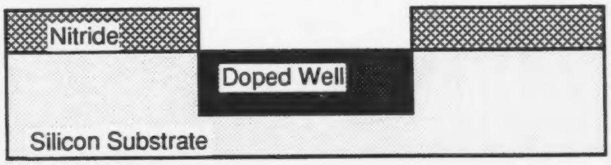

Nitride
Silicon Substrate

Figure 2.14 Etching process steps 3 to 6 for high energy ion implantation

Step 7 - High Energy Implant


Nitride
Silicon Substrate

Step 8 - Well Drive


Nitride
Doped Well
Silicon Substrate

Step 9 - Nitride Etch

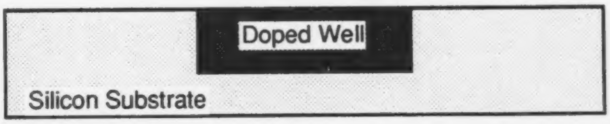

Doped Well
Silicon Substrate

Figure 2.15 Etching process steps 7 to 9 for high energy ion implantation

Through use of ion implantation and diffusion processes, the basic semiconductor conduction characteristics and structure is defined. The process described above is a basic overview of the processes. These n type and p type semiconductors form the base and emitter regions in bipolar technology and form the drain and source regions in MOS technology.

2.3.5 Overall Processing

Construction of a working circuit involves using elements of the four main areas of semiconductor processing - ion implantation and diffusion, photolithography, etching and thin film deposition - to build a circuit layer by layer. The basic steps are repeated with different materials within the process. For example, the defining of the metal lines is done first with the deposition of the metal. The desired interconnect pattern - the "lines" are exposed onto photoresist that is on top of the metal. The photoresist is then developed, exposing the metal where there will be spaces between the metal lines. This is etched away then the photoresist is etched. All that is left are the metal interconnect lines. The cycle is then repeated with different materials for the next process. An insulator such as oxide may be deposited and masked. Interconnect holes are then etched in this oxide. Further process steps then occur to build the circuitry, layer by layer. The basic types of processes are repeated over and over in the construction of the circuit. The materials used and the exact processes change, depending on the layer and on the type of semiconductor circuit being constructed. Even though there are exceptions, most of the semiconductor processes used fall into one of the four basic groups.

To fully understand the processing technology, a hypothetical ASIC array is constructed on the next few pages. The circuit in question uses one logic cell to make two invertors. The process described is only an overview of how this gate array is customized. All the layers and processes are not described. The technology illustrated is that of CMOS dual layer metal gate arrays. The next pages containing Figures 2.17 to 2.22 will illustrate the construction of the layout of a gate array layer by layer viewed from the top. The structures and layouts used are similar to actual ASIC gate arrays constructed by LSI Logic corporation. The diagrams illustrate the construction, from the base array through to the completion of the pad etch step. A diagram of a 3 layer metallized CMOS cross section is shown as an illustration in Figure 2.16 to provide some insight into the 2 layer metallized topological views shown in Figures 2.17 to 2.22.

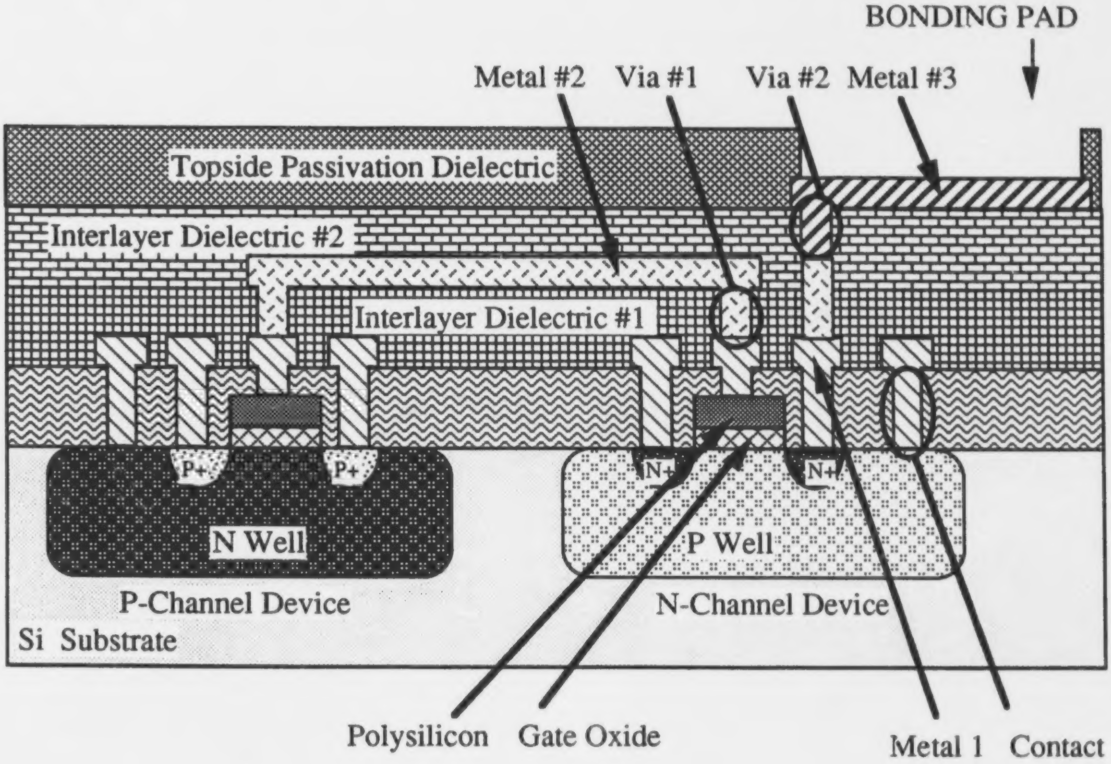

BONDING PAD
Metal #2
Via #1
Via #2
Metal #3
Topside Passivation Dielectric
Interlayer Dielectric #2
Interlayer Dielectric #1
P+
P+
N+
N+
N Well
P Well
P-Channel Device
N-Channel Device
Si Substrate
Polysilicon
Gate Oxide
Metal 1
Contact

Figure 2.16 Cross section of a 3 layer metallized CMOS device

The Figure 2.17 shows the basic CMOS cell. Figure 2.18 shows the circuit with the contact holes defined. Figure 2.19 shows the first metal layer routing. Here the pads are defined as well as the power bars that supply the cells. Although the example only shows a single cell, the basic structures are repeated again and again in actual gate arrays. Figure 2.20 shows the second metal as in the real gate arrays. Figure 2.21 shows how the via holes are defined. Lastly, Figure 2.22 shows the pad openings in the protective passivation.

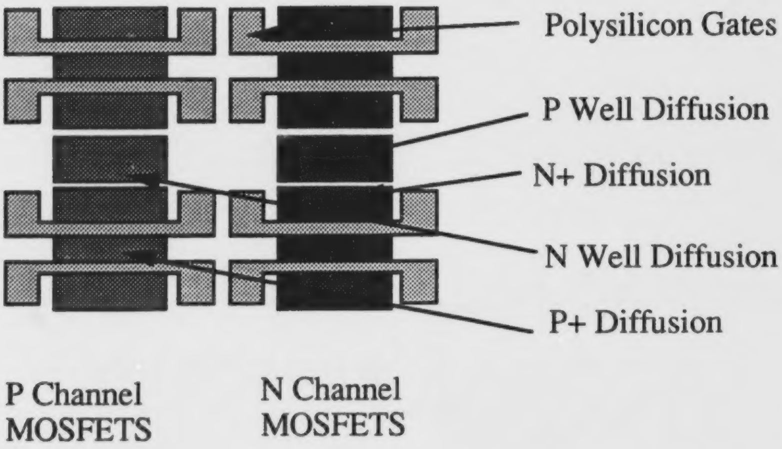

Polysilicon Gates
P Well Diffusion
N+ Diffusion
N Well Diffusion
P+ Diffusion
P Channel MOSFETS
N Channel MOSFETS

Figure 2.17 Basic CMOS cell - base array - one cell

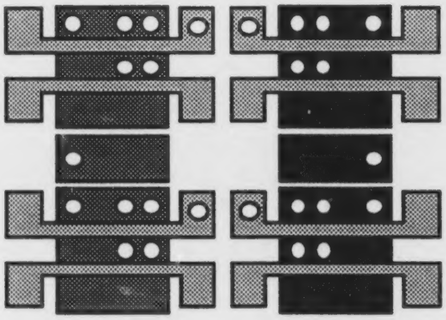


Figure 2.18 Basic CMOS cell - contact holes defined

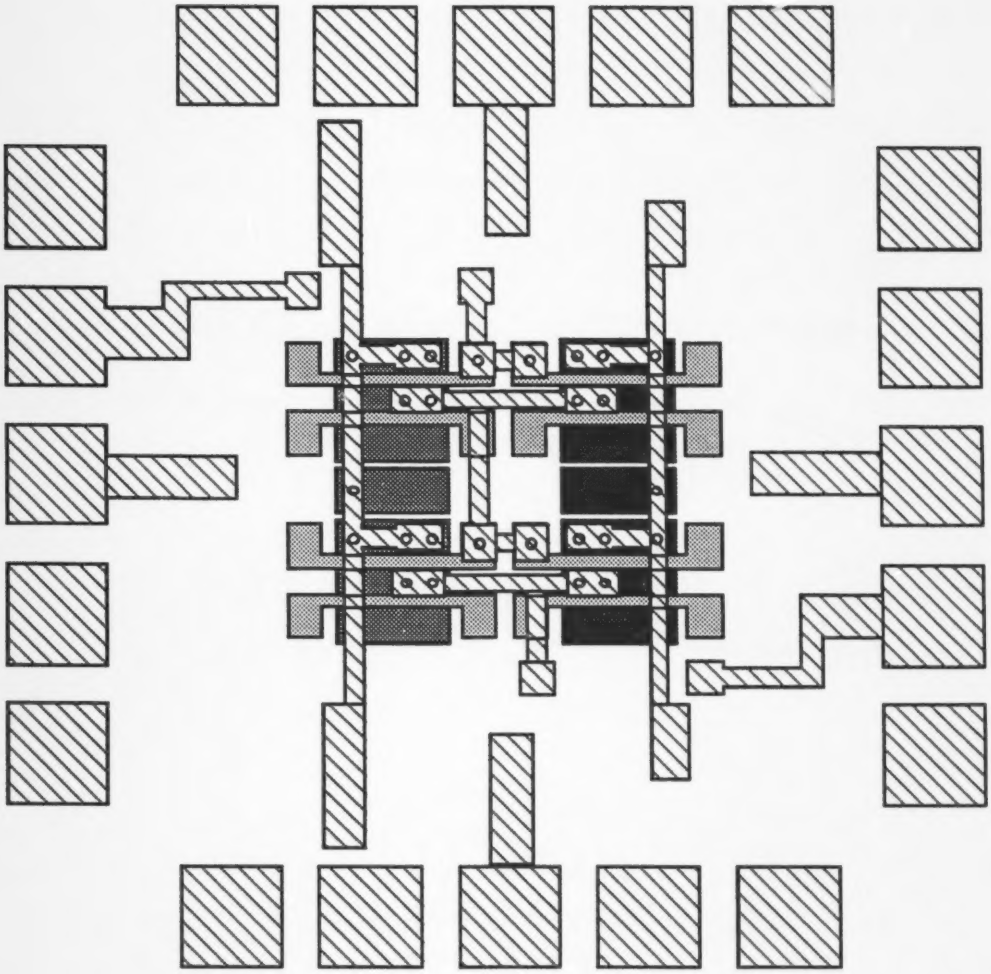


Figure 2.19 Basic CMOS cell - first metal layer routing

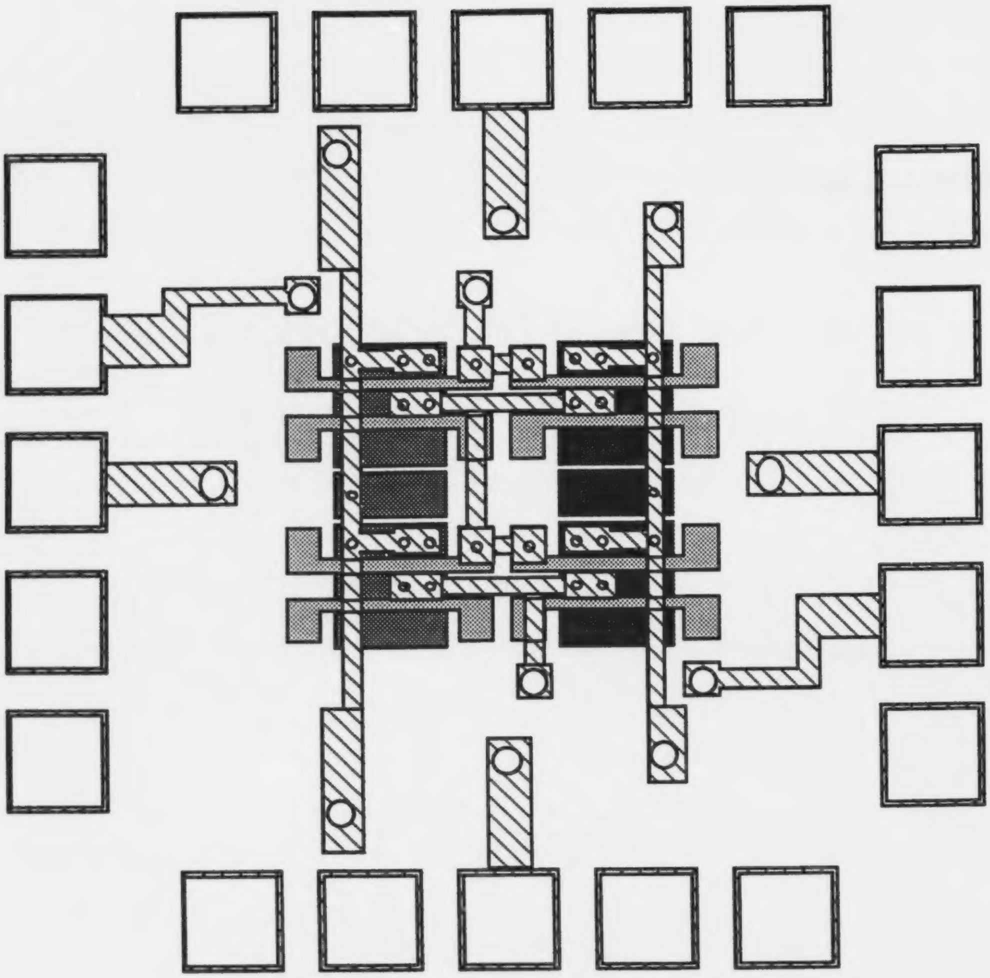


Figure 2.20 Basic CMOS cell - via holes defined

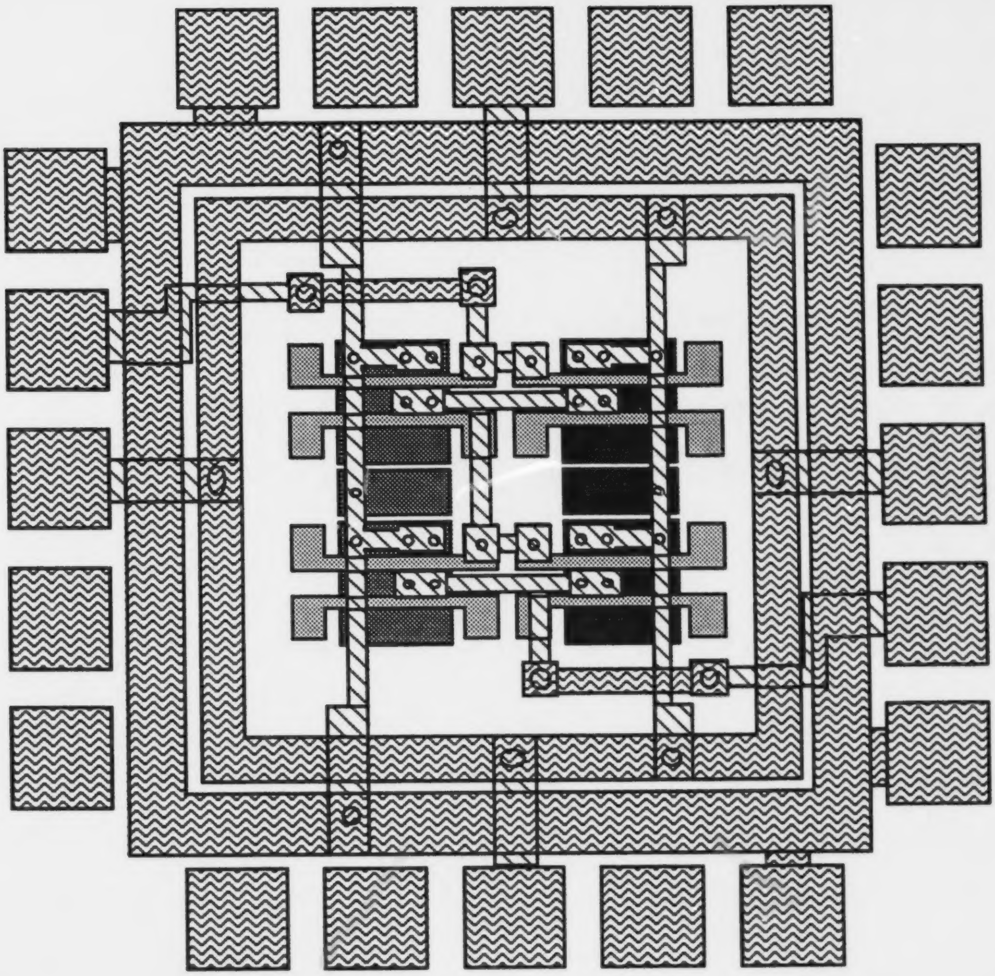


Figure 2.21 Basic CMOS cell - second metal layer routing

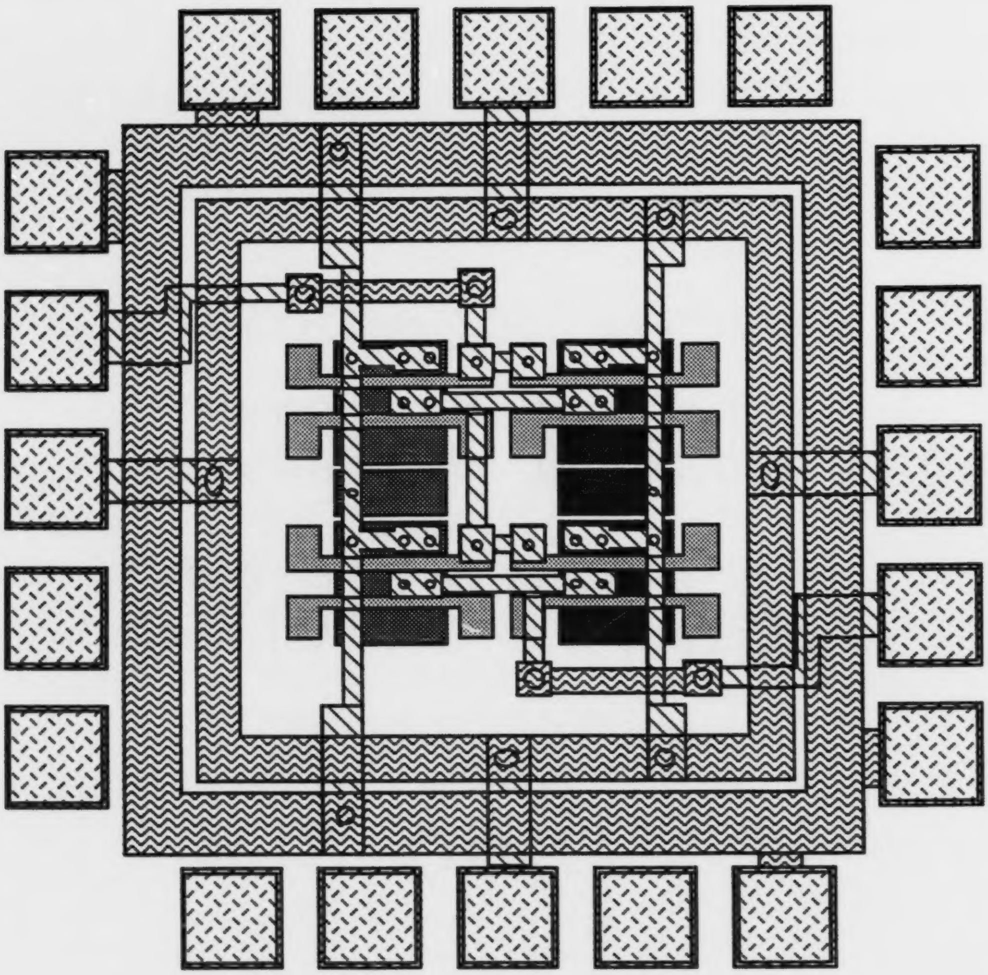


Figure 2.22 Basic CMOS cell - pad opening in the protective passivation.

2.4 ASIC Logic Circuit Basic Operation

The gate arrays illustrated previously are made from CMOS technology. CMOS stands for Complementary Metal Oxide Silicon transistors. The complementary refers to the fact that there are two transistors that work in tandem. The two MOS transistors are of two types - N channel and P channel. The N and P are the types of silicons as illustrated earlier. The transistors have been doped through ion implantation and diffusion such that they will act as switches. They will effectively turn on and off with changes in voltage. Turning off is when the transistor effectively stops conducting. The transistors act in tandem by having the N channel transistor turn off, i.e. stop conducting while the P-channel transistor turns on. This is effectively how the two transistors work together - in a push- pull type of arrangement.

A MOS (Metal-Oxide-Silicon) transistor has four terminals. They are: gate, drain, source and substrate. The gate is the control. Depending upon the voltage on the gate, current will flow between the source and the drain. The substrate is used for biasing. A diagram of a CMOS cross section is shown in Figure 2.23.

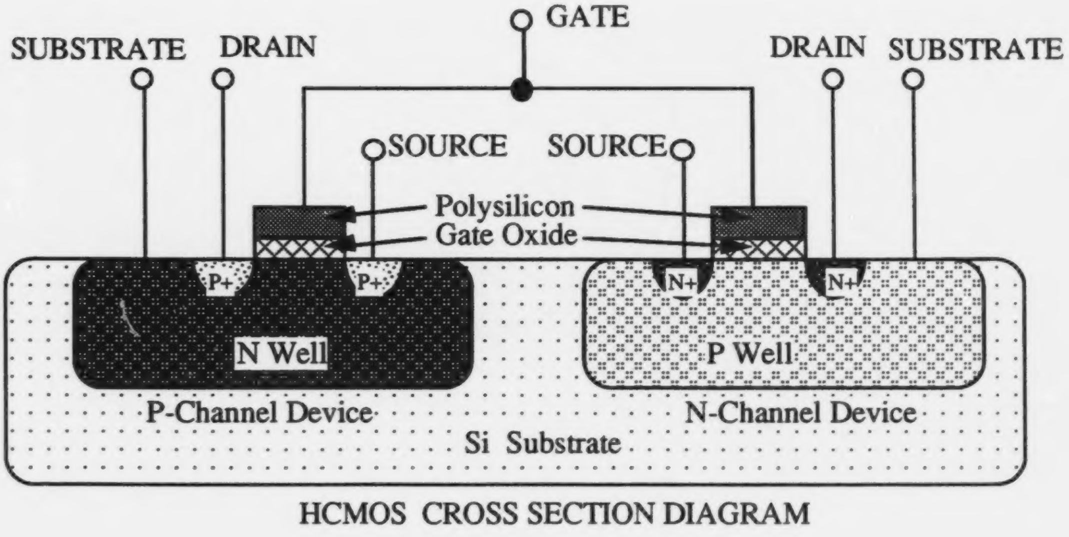

GATE
SUBSTRATE
DRAIN
DRAIN
SUBSTRATE
SOURCE
SOURCE
Polysilicon
Gate Oxide
P+
P+
N+
N+
N Well
P Well
P-Channel Device
N-Channel Device
Si Substrate
HCMOS CROSS SECTION DIAGRAM

Figure 2.23 Cross section of CMOS transistors

For the purposes of understanding yield and reliability, a basic description of how a MOS transistor works is as follows: In the diagram of the cross section, the gate is constructed of polysilicon on top of the gate oxide on top of the p-well. Both the p-well and the polysilicon have been doped to be conductive. The gate oxide is constructed of highly pure silicon dioxide. This is essentially a capacitor. The p-well is p type silicon. This means that it has more holes or places for capacitors in its crystalline structure. The source and drain area are n type silicon. There is an excess of electrons in the crystalline structure. The N-channel transistor is usually powered by a five volt power supply.

To explain how the N-channel transistor works, the source and substrate are connected to ground, or 0 volts. The drain is connected to five volts. The n-p-n structure between the source and the drain has a reversed bias diode in the structure. There is essentially very little or no current flow between the source and the drain. Now, if the gate of the transistor has been at 0 volts, this condition will remain. This is shown in Figure 2.24.

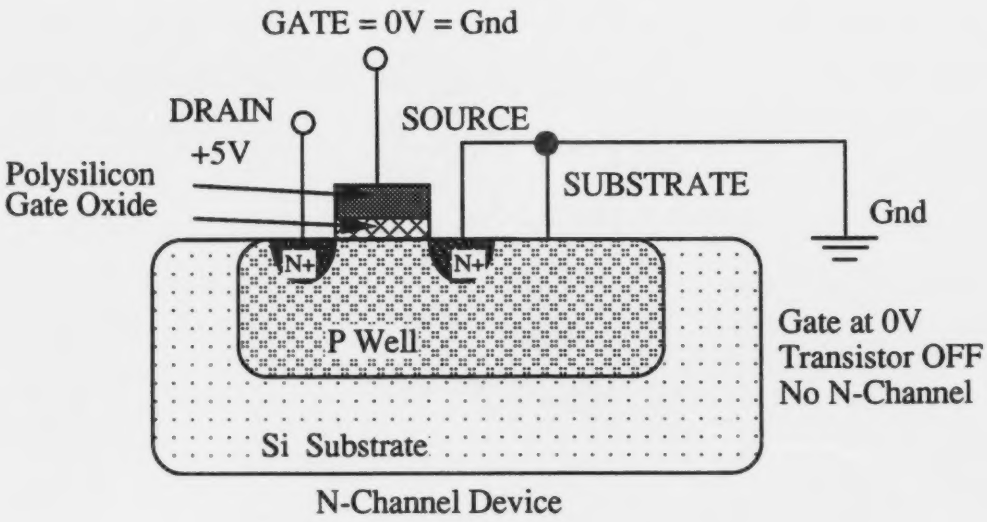


Figure 2.24 N-Channel transistor with gate voltage at 0 volts

If a positive voltage is put on the gate, the following happens. Even though there is a p-well with an excess of holes, i.e. more spaces for electrons than normal in the crystalline structure, there are still electrons. The positive charge applied to the gate will start to attract the electrons to the top of the p-well layer, just under the gate oxide. The electrons will come from the zero volt n+ source region. Therefore, there is a region formed near the source and under part of the gate where there are an excess of electrons present. This is shown in Figure 2.25.

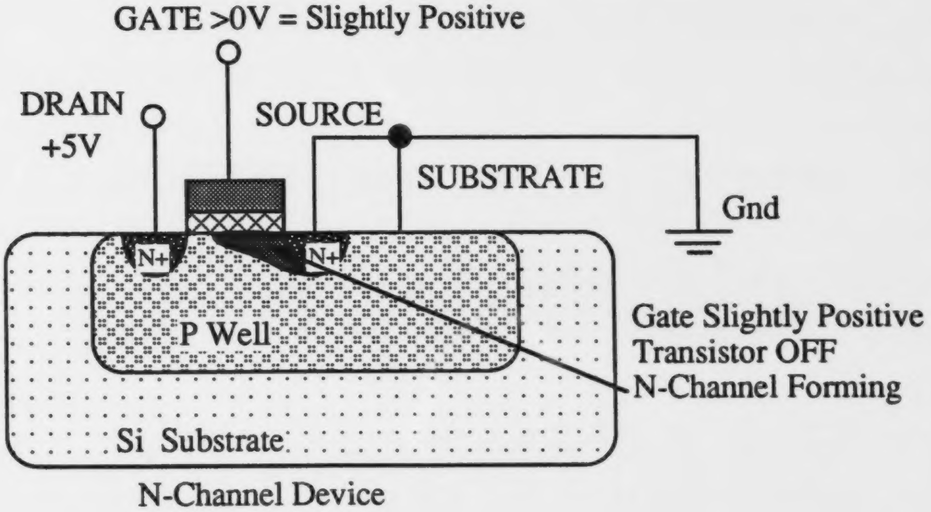

GATE >0V = Slightly Positive
DRAIN
+5V
SOURCE
SUBSTRATE
Gnd
N+
N+
P Well
Si Substrate
Gate Slightly Positive
Transistor OFF
N-Channel Forming
N-Channel Device

Figure 2.25 N-Channel transistor with gate voltage > 0 volts

This region will grow as the voltage on the polysilicon-oxide-p-well capacitor is increased. As more charge or voltage is applied, this region of electrons will start to grow until it extends all the way from the source to the drain. Now this region has excess electrons so it is like a N type silicon region. When the region grows to the point where it is entirely between the drain and the source, there is essentially a n type silicon between two n type silicon regions. This region is called the "N Channel" Since this N channel is of a similar type material as the source and the drain, the resistance is low, that is, current will now flow between the source and the drain . The transistor will turn on. By applying charge to the gate, a N channel is created between two N type semiconductors and the transistor will turn on and start to conduct. The typical voltage that this occurs at is around 0.7 volts. This is illustrated in Figure 2.26.

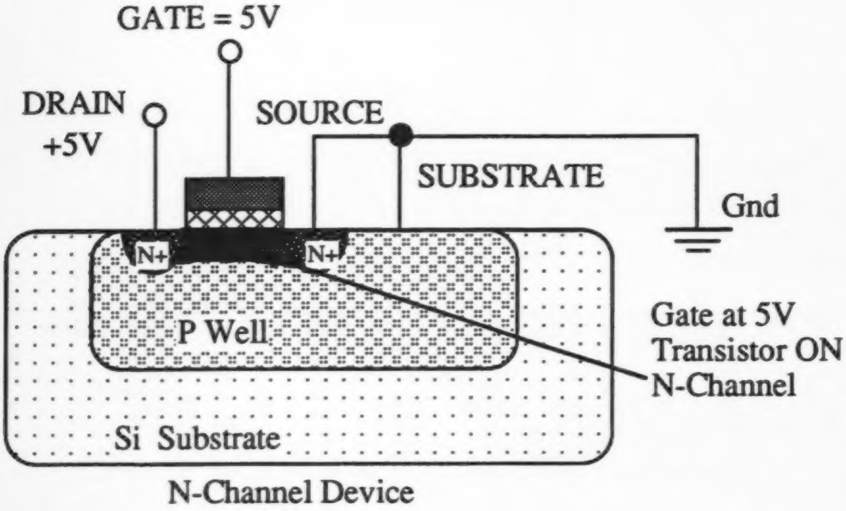

GATE = 5V
DRAIN
+5V
SOURCE
SUBSTRATE
Gnd
N+
N+
P Well
Gate at 5V
Transistor ON
N-Channel
Si Substrate
N-Channel Device

Figure 2.26 Fully on N-Channel transistor

For the P channel device, the similar effect occurs but the charges are reversed. The gate is biased lower or more negatively than the substrate. This means that electrons will be forced away from the region near the source and the gate oxide. A P channel will start to form. The drain is biased more negatively than the source. Conduction will occur when the p channel forms under the gate oxide between the source and the drain. The transistor turns on. If the substrate and the source are connected to five volts and the drain grounded at zero volts, the transistor will turn on as the gate voltage is decreased from five volts. Biasing voltages of this type is common in CMOS circuitry. The P channel transistor will turn off as the N channel transistor turns on.

CMOS circuitry takes advantage of how the two transistors work in compliment. If we take the basic invertor of our example chip, the schematic diagram is shown in Figure 2.27.

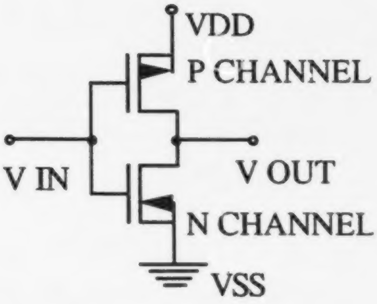


Figure 2.27 Basic invertor circuit

The invertor is the most basic of the logic circuits. An invertor is a logic circuit that the output is opposite for the input. This means that if the input was a logic 1 or five volts, the output would be at zero volts or a logic zero. Similarly, if the input is zero volts, the output is at five volts. To see how the CMOS invertor works, the assumption made is that the transistors work as switches. This is a good approximation for CMOS circuits since the doping and the transistors are designed such that the transistors have high source drain resistance when off, low when on, and have quick switching characteristics. If the gates are driven to a logic 1 or five volts, the N-Channel transistor will be on and the P-Channel transistor will be off. The output is then essentially connected to zero, or ground. If the input is at zero volts, the output will then be five volts because the N-channel transistor will be off and the P-channel will be on, essentially connecting the P-channel to the 5 volt supply.

One of the main reasons that the use of CMOS logic circuits is so popular is that the circuits use little power. If we again look at the sample of our basic double invertor cell as shown below, we will see two invertors connected together, as drawn in the circuit diagram shown in Figure 2.28.

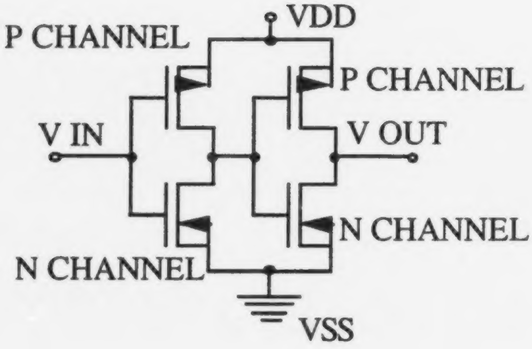
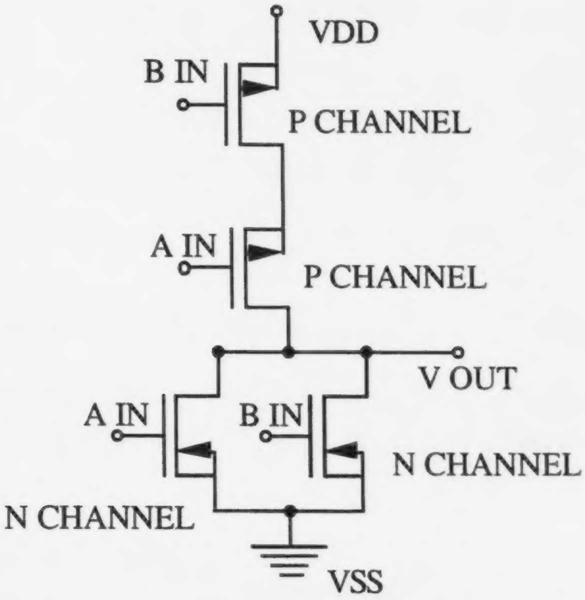


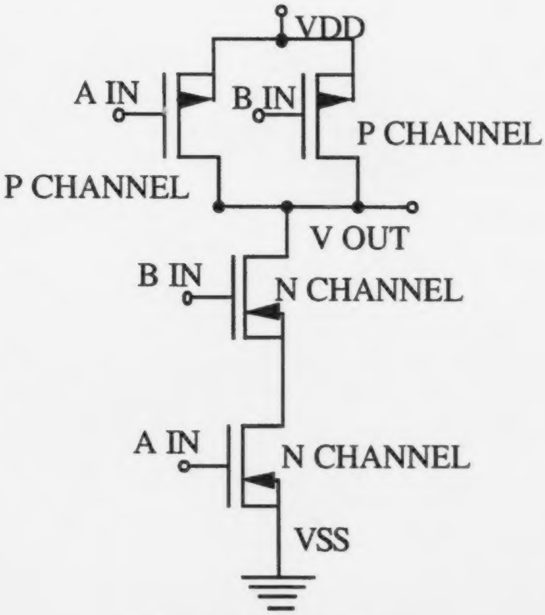
Figure 2.28 Double invertor circuit diagram

The output of the first invertor is connected to the input of the second. The input of a CMOS circuit can be approximated by both the capacitance of the N-channel gate and the P-channel gate. As the output of the first invertor switches, the gates of the second invertor are either charged or discharged. After a finite time, the gates of the second invertor reach their steady state. Since this is like a capacitor, current essentially stops following. The only current present is the gate leakage current, which is usually in the order of nanoamps. This means that larger current only flows when the gate conditions change; i.e. when transistors are switching and the capacitance of the gates are charging. After a time, the circuit charge changes propagate through and the chip settles into an idle state. In this ideal state, very little current is drawn.

By using more than one transistor in the circuit, more complex gates can be made. For example, if there are two inputs A and B in two different configurations as shown in Figure 2.29, there are two more of the basic logic gates formed: NAND and NOR logic gates.


VDD
B IN
P CHANNEL
A IN
P CHANNEL
V OUT
A IN
B IN
N CHANNEL
N CHANNEL
VSS

CMOS NOR GATE


VDD
A IN
B IN
P CHANNEL
P CHANNEL
V OUT
B IN
N CHANNEL
A IN
N CHANNEL
VSS

CMOS NAND GATE

Figure 2.29 CMOS basic logic gates

By using more transistors in the cell and multiple cells, more complex circuits can be formed. Interconnection of these circuits form microprocessors, etc. The basic cell structure can also be adapted to act as memory. By using the capacitance of the gate themselves, a charge can be stored, that is, it can be remembered.

The basic understanding of the processes and the structure of the circuits, cells and integrated circuits is required to understand how these circuits are affected. Understanding of how the circuits work and how they are built is necessary to understanding what goes wrong when they do not work.

CHAPTER III
YIELD AND RELIABILITY FUNDAMENTALS

3.1 Introduction

Semiconductor designs and processing have undergone rapid changes in complexity, performance and size. Along with this increase in technologies, there have been increases in economic pressures on this highly competitive industry. What was once a small industry located almost entirely in Silicon Valley California, is now spread through out the world. With the strong change towards quality and performance, brought about largely by the Japanese manufacturers, the industry is more competitive than ever. To survive as a manufacturer, one must produce parts that are cheaper, more reliable, and of higher performance than ever before. Obviously, with the increasing size and complexity of circuits, the increasing complexity of manufacturing, coupled with the decreasing geometries the question arises: How can any manufacturer continue to produce parts cheaply and reliably? The answer lies in more quality built into the manufacturing along with the increasing monitoring and testing of both yield and reliability.

Quality is now being manufactured into the processes in most wafer fabs. This is done through the application of techniques such as SPC - Statistical Process Control and DOE - Design of Experiments. The goal is to fully understand and characterize processes and to detect processes that are starting to go out of control before they actually are out of control. Process windows are the process region that the processes can vary within and still not effect the desired outcome on the product. The processes used in IC manufacturing are designed to have the widest processing window as possible. All processes have natural variations. For example, a plasma etch step may have a large pressure process window. If there are changes in the pressure of the plasma etch machine, within the pressure process window, the same desired etching result is achieved on the product. The process can then be centred in the middle of this process window. The natural variation of the process is usually much smaller than this window. The process is then what is termed capable. If there are changes in pressure outside of the window, then the desired etching result will not be achieved.

Manufacturing philosophy has changed. Instead of being **reactive** - that is reacting to processes and conditions only when they are out of control - manufacturing must be **proactive** - designing in quality and reacting before processes and conditions are out of control. In the semiconductor industry, this is the approach that must be taken in order to remain competitive. The driving force behind all of this is to stay in business and remain profitable. It is to produce higher quality parts cheaper and faster than your competitor. With the increase in quality production, there has been increases in the yield. The yield is the number of good parts that has completed processing. Yield is a measurement of the manufacturing performance. Within manufacturing plants, the term "yield" is used to describe the performance of various steps within the manufacturing. There are various "yields" taken - line yield, wafer sort yield, assembly yield, final sort yield, burn in yield, etc. Line yield refers to how many wafers make it through the manufacturing process. Wafer sort yield is the number of good die that the tester has determined that are good on a particular wafer. Assembly yield refers to the number of die that have been assembled into packages correctly. Final test yield is the number of good chips that the tester has determined to be good. And finally, burn in yield refers to the number of good chips that the tester determines are good after the chips have been exposed to an infant mortality burn in.

The actual terminology and "yields" will vary from manufacturer to manufacturer. However, the types of yields listed above will be present in most manufacturers today. The most critical yield that is used as a measure of performance by most wafer manufacturers is the wafer sort yield. This is the first time that the die on the wafer are subject to a complete functional electrical evaluation. The good die are die that pass the compliance test on the tester. That is, there is a test program that the tester executes. The DUT -Device under test (the die) - is then subjected to these various tests, as described by the test program. All the die on a wafer are probed and tested for complete functionality. Traditionally, this is where most of the yield loss of potentially good parts occur.

Yields are critical economically. The cost of producing a wafer depends on the costs of running the wafer through the wafer fabrication processes. It essentially costs the same amount of money to produce a good wafer as it does to produce a wafer with no good die on it. When orders are placed, they are placed for the amount of good working ICs. The more good working die there are on a wafer, the fewer wafers it takes to meet a particular order. This has an immediate two fold effect on the economics involved. Firstly, the cost on a per die bases goes down. The secondary affect is an increase in capacity. Wafer fabs have a fixed capacity of being able to produce a certain number of wafers per week. If it requires less wafers to meet a particular order, this then frees up other wafers for additional orders and customers. This means an increase in revenue generated by the wafer fab without an increase in cost.

Yields are only part of the picture. The parts must work when put into their particular application and continue to work. The final product must also be reliable. Again, the approach to guaranteeing this reliability is to be proactive. Reliability should be designed and tested into the manufacturing processes. As the yields and processes are constantly being monitored for out of control and abnormal occurrences, so must the reliability of the chips be similarly checked. The issue again comes down to economics. When a part fails due to the manufacturing process, the part is sold back to the manufacturer. If the customer has not gone with another supplier, he must then occur losses in his production while he waits for good, reliable parts.

Traditionally, within semiconductor manufacturers, yield and reliability were treated as separate areas within the manufacturing processes. With the increasing demands upon manufacturing and the resulting changes, there is a large overlap between the two areas. To understand these changes, one first must have a basic understanding of the reasons why parts fail. In other words, a basic understanding of the defects caused by manufacturing is important to understanding the mechanisms that affect both yield and reliability.

3.2 Defects

A defect is something that is undesirable on a die. This appears to be a very large open ended definition and it is for a very good reason. If one were to attempt to define all the variables that affect manufacturing of Integrated Circuits, one may as well attempt to count the stars in the sky. And with each of these variables, there can be numerous factors that cause failure on ICs. failure on wafers have been reported to occur by such obscure events such as a result of a certain perfume worn by female operators. Why is this so? The complexity and the amount of processing required to make these high technology chips is ever increasing. As the complexity increases, so does the sensitivity to defects of all kinds.

Defects can be classified. The first major classification is into killing and non-killing defects. A killing defect is a defect that prevents the desired functionality of the circuit in question. A non-killing defect is a defect that does not cause circuit failure but in other cases, may cause circuit failure. This is best illustrated through the use of an example. Metals such as aluminium are used in semiconductor manufacturing for interconnection. There are instances when a particle occurs between two metal lines. There may be some metal that has been left around the particle in question due to the process. These two lines are shorted together causing the circuit to fail. This is an example of a killing defect, and is shown in Figure 3.1. This same type of defect can also be a non-killing defect. If the same particle happened to land in a different spot on the die, as shown in Figure 3.2, it would not cause circuit failure. In Figure 3.2, even though there is still metal around the particle, there is no shorting of circuitry. The particle is essentially imbedded in the circuit in an area where it can cause no harm. It is therefore a non-killing defect.

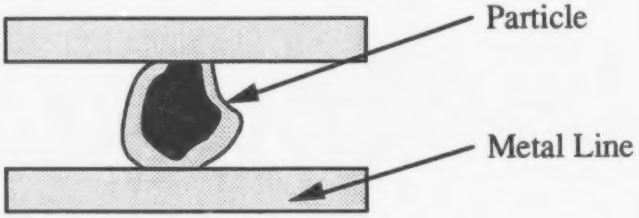


Figure 3.1 Killing defect

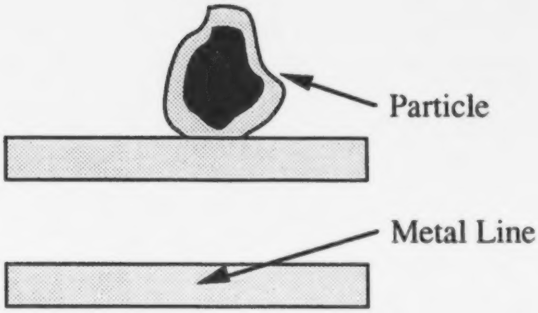


Figure 3.2 Non-killing defect

The particular defect illustrated in both Figures 3.1 and 3.2 can come from multiple causes. This defect may be a result of a particle landing on the wafer during the metal thin film deposition step. This may even come from the sputtering deposition system that may be used for this particular process step.

The defect could even be caused in the photolithography area. As discussed before, photoresist is spun onto wafers. If a particle were on the photoresist, then the defect shown could appear on a single die. If the particle were on the mask, then multiple die would have the same defect in the same location. Multiple die with defects are possible if a stepper were used for exposure, as discussed in the photolithography section. If the mask field exposed two die at a single time and the defect was on the metal 1 mask field, then one half of the die on the wafer would have this defect. Remember that a single layer field can consist of more than one die. This type of defect is called a repeating defect. It is the most damaging defect of all to yield because one small defect is repeated over all or a majority of the wafer. The last place in the process where the metal defect particle could take place is in the etching process. The particle could land on the wafer just before etching and give the result seen in either Figure 3.1 or 3.2. It is very important to have a basic understanding of the processes a wafer goes through during manufacturing in order to identify the potential sources of killing defects. Even with a basic understanding of the process, further tests on the production line may be required to narrow the defect source down to a single process or machine.

A killing defect may not be the killing defect that has caused the die in question to fail.

There can be more than one killing defect on a bad die. The defect that caused the die to fail is the defect that the tester program first detects. This means that a killing defect can cause a functionality failure in one part of a die before another part of the die is tested. If the test were continued, other killing defects would be detected. However, it is not economically feasible to test all the parts 100%. Once a failure to comply with the required test program occurs, the testing stops on that particular die and starts on the next die. The die is declared "bad" by the tester. The reason that it is not economically feasible to continue testing on all the bad die is that continuing the test wastes tester time. This reduces tester capacity. It takes a finite time to test each die. Even though you want to test to completion on the good die, you want to determine that a die is bad as quickly as possible. By doing this, the time the tester takes on bad die is minimized. This results in more time available on a particular tester for good die, thus increasing the throughput for good die. For example, if a wafer has 200 die on it and each die takes 30 seconds to test, then the total test time for a wafer with 100% good die would be 100 minutes. If there were 20 bad die and the test time of the bad die was an average of 10 seconds each, the total time on the tester is now approximately 94 minutes. The extra 6 minutes hopefully can be used to test more good die.

The example of the metal bridging particle is also an example of a point defect that can be due to contamination. The particle can be a result of equipment in the process, the environment that the wafer went through in the process, an operator passing too near the die, residue from cleaning, residue from a scratch elsewhere, etc. Anything the wafer "sees" throughout the wafer fab processing can be a source of contamination. Contamination is usually classified as point defects because they affect a single point within a die (usually). Global defects are defects that tend to affect entire wafers. An example of a global defect is having no via holes between the first metal and the second metal interconnection layer. This could result by having a failure during the via etch step. A possible failure such as this would be operator error. The wafers were sent on to the next process step without the via etching occurring, resulting in no vias. The entire wafer was effected globally.

Usually, there are procedures in place that detect the kind of defects described above while the wafer is still in processing within the wafer fab. The wafer is then taken out of production and declared of no value. The wafer sort yield is not affected in this case but the line yield is. There is one less wafer that will come out of the wafer fab than was started as a result of this. It is important to detect defects early in the process, if possible. By continuing to process the wafer, it costs money. If a bad wafer can be detected and declared "dead", then there is none of the cost incurred in finishing that wafer. This extra cost would be worthless for no die would yield on the entire wafer in the case of the missing via example.

Global defects tend to be a result of some problems or mistakes within the process. Wafer fabs that are considered "mature" tend to have fewer global defects than point defects. A "mature" process would have identified most defects that are caused by the process and have eliminated them. There would also be controls and checks in place to prevent and detect most global defects that affect wafer sort yield.

3.2.1 Defect Densities

Defects can occur in a myriad of sizes. Every wafer has defects on them. It is the size of the defects and the process sensitivity that are important. Defects can kill die from the atomic level on up in size to dust particles hundred of microns large. The size is dependent on the type of defect. A approximate rule of thumb can be used when it comes to particulate defects. A particulate defect that is one tenth of the minimum sized feature will generally not be a yield or reliability problem. Of course the larger the defect size, the more chance there is that the defect will occur in an area of the die where it will be a killing defect.

The concentration of defects is referred to as the defect density. The more defects there are, the greater the chance of the defect being a killing defect. Defect density is defined in number of defects per area. The defect density is usually given as an average number of defects per area.

3.3 Yield Modeling

Defects are important to be able to predict yields in semiconductor wafer fabrication plants. Predicting yields of similar products is important, especially in the ASIC business (Application Specific Integrated Circuit). When running a particular chip for the first time, some knowledge of the expected yield is important. By having an expected yield, the correct amount of wafers can be started in the factory. If too many wafers are started, then there are too many parts. This increases the cost of each part that is sold. If too few are started, then the order ends up short with shipment delays. Yield modeling helps to predict what the yield should be. This is essential to factory planning and to correctly loading the factory. Yield modeling also is an important tool in analysis. If the actual yields are quite different from the predicted yields, then this particular data may provide clues for improvement to the process, design procedures and even the modeling assumptions.

All yield modeling is based on a ratio of good die to bad die on the wafer. Generally, most yield models are of the form [8]:

$$Y = Y_0 Y_1 (D_0, A, \beta) \qquad (3.1)$$

where:

Y = Yield in % of total die
Y_0 = The fraction of sites that do not have process related defects or circuit sensitivities (global defects)
Y_1 = Function that describes the yield on these potentially good sites
D_0 = Defect Density Function (point defects)
A = Area of the Chip
β = parameters unique to different yield models.

Most of the models in popular use throughout the industry are based on the type of model shown above. There are assumptions made in determining the different defect density distributions. Different yield models are based on different killing defect density distributions. All yield models are based on the probability of finding no killing defects in a given area, assuming a certain defect density function. Two of the most popular models are given below [3]:

Poisson Model:

$$Y = Y_0 e^{-DA} \qquad (3.2)$$

Murphy Seeds Model:

$$Y=Y_0e^{-\sqrt{DA}} \qquad (3.3)$$

What is important to remember about yield modeling is not the modeling itself but the application of the modeling. No yield modeling has been developed that has been proven to predict yields absolutely. Yield modeling is used as a guide line and more for comparison of yields of chips of different sizes. On average, they are a good guide line but the natural variability of a process makes it impossible to predict all the individual wafer yields.

What most yield models show is a dependence upon both the defect density and the area. There is usually a geometric dependence upon the area. As the die get larger, the yields tend to decrease geometrically. While it is possible to be yielding in the high 90% range on small and very small die, a good yield is over 50% on larger die. As the complexities of dies have increased, it has become more and more difficult to yield. It is also harder to yield on these larger die that are now demanded by the customers. To keep costs down, these larger die now contain smaller and smaller geometries thus making them more sensitive to defects. The routing density of these die are also increasing. This means more metal interconnection per given area. Design software and CAD tools have made it easier and possible to design and simulate more complicated and denser circuits. The demands on processing and quality of processing have increased dramatically. With the competition in the industry and the demands of the designers, the increased manufacturing and performance demands show no sign of slowing down.

A comparison of the Murphy Seeds model and the Poission model with actual data is shown in Figure 3.3.

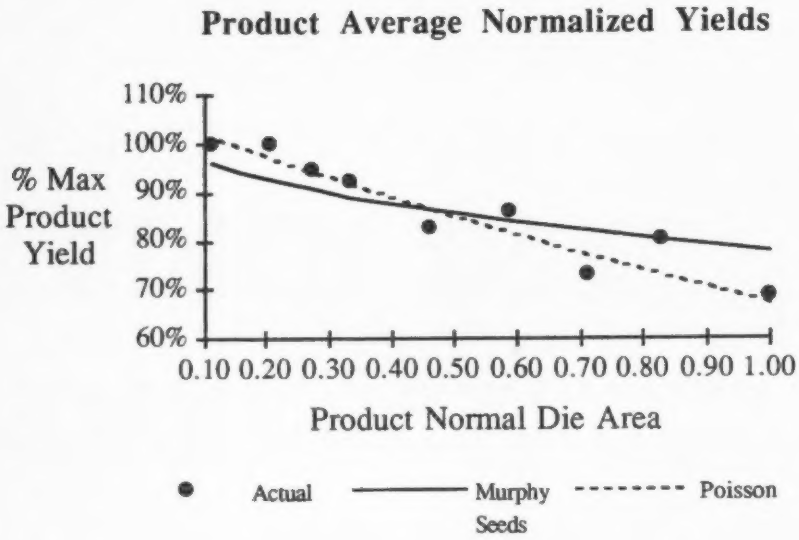


Figure 3.3 Comparison of yield models

3.4 ASIC Logic Circuits Test Methodology

Another issue in looking at yields is the testing for defects and for circuit functionality. As explained previously, testing is also subjected to the practical constraints of economics. It is easy to forget that the tester, the test program, the probe card and the interconnections have an effect on the result of the test as well as the device under test. If there are problems in any of these areas, the resulting test results can be false. A device that is good may be declared bad. The flip side of this is that if there is a problem in the test program, devices that are bad may be declared good. This is a serious issue. Customers are not happy with vendors when they receive parts that do not work. The other downside is that if good parts are not found, there is a loss of revenue.

Most test methodologies involve a hierarchical test structure. The structure refers to the fact that the tests are done on a circuit are in a particular order for a very good reason. The first tests done are intended to be the most basic types of test and are intended to produce the first failures. Most ASIC test methodologies have some similarities to the following order of tests:

1. Continuity Tests
2. Power up Test
3. Basic Functionality test
4. Fine Functionality, Current and Specific (Parametric) tests.

The continuity tests test both the Device Under Test (DUT) and the tester. It is a test to first determine that the probe card is making contact with the device, and that there are no faults with the testing connections from the tester to the device. The test may also check for both open and short conditions. This test will also detect problems on the die that can cause opens and shorts. Generally, most of the bonding pads will be checked during this test.

After good connection with the part is established as a result of the continuity tests, the first real test of the circuit internal workings is done. The part is powered up and checked to see if it is not drawing significantly more than normal operating current. If it is, this usually means that there are defects on the circuit that cause a high current draw.

The last of the first simple tests is the basic functionality tests. The circuit under test (DUT) is excited with specific inputs. The appropriate output conditions are monitored. This is then repeated for the next functional blocks, if the first functional block passes. The test continues until the circuits entire functionality has been tested for. Particles and other contamination can be a typical cause of failing the basic functionality tests.

These first three tests tend to detect more than 90% of the failures found on circuits. The remaining tests are designed to test the part for leakage and other DC parametric performance. Speed testing is usually done here. In a mature process, there usually is no more than 10% of the failures in this part of the test methodology. The first two tests - continuity and power up- usually do not take up significant tester time. The design of this methodology is to weed out the worst circuits first and spend time just testing the potentially good parts.

The tester decides whether or not a part passes by comparing the output it receives back from the part under test to the limits defined in its test program. The only time that test program problems are a large issue is usually when a particular chip is run for the first time in the factory. If a part fails, failure analysis must be done to determine if the part has failed for one of the possible three major reasons: the part's design, defects on the part caused by the manufacturing of the part in the wafer fab, or failure in the testing of the part. Once a part has been shown to yield successfully, the design and the test program is usually not a factor, but the tester may still be a factor.

Most testers now consist of sophisticated computer controlled power supplies, relays, and meters. They require constant attention and calibration. The wafer probers that bring the die in contact with the probe card can also be a source of problem. Generally, mature manufacturing sites will have procedures such as calibration, preventive maintenance and checks on the testing to catch erroneous testing results.

3.5 Wafer Level Reliability Testing

In semiconductor manufacturing, the yield and reliability of the parts have been traditionally treated as separate issues and departments. It was believed that the defects that caused yield problems could be detected at the wafer sort step and the reliability of the parts could be tested for by burn in and life time testing procedures. The feedback was then to the engineers involved with the processing set up and control. The processing engineers traditionally tended to be more concerned with the immediate yield concerns and keeping their processes in control. Reliability tasks traditionally involved initial burn in of finished parts to detect and weed out infant mortality failures and life time testing though the application of stress to finished ICs. The ICs tended to be sampled from lots. The life time tests results were not known for a period of months because the traditional type of lifetime testing involved powering up the ICs at a higher than normal operating temperature for a period of time.

Traditionally, testing for product lifetime meant a life test through the use of five common types of stressing on the parts - temperature, voltage, current, humidity and temperature cycling [7]. Most product lifetimes are determined through temperature acceleration, based on the Arrhenius Equation [4]. Parts are "burned in" to determine their lifetimes. Burn in refers to placing packaged parts in a burn in oven, powering the parts up, and periodically testing the parts.

Upon failure, analysis is done to determine the defects that caused the failure. The procedure of burn-in and analysis usually takes a period of months to complete. Most semiconductor manufacturers today adhere to an industry accepted military Standard MIL-STd-883.[9]. This involves a 1000 hour operating life test at 125℃. The problem is that the time involved here is a period of months. With the complex processing that is going on in todays semiconductor manufacturers, a lot of bad, unreliable product can be produced during this period. The feedback from a problem could take months.

One of the biggest factors influencing the reliability picture is that customers are demanding a higher quality IC than ever before and higher assurances of quality but want a lower cost IC than ever before. To insure higher levels of quality through testing and yet continue to reduce the cost of the final product is an almost impossible task. To sample more ICs for the traditional way of reliability testing costs more. The answer? The semiconductor industry is working on incorporating more Wafer Level Reliability testing into the manufacturing processes. Wafer Level Reliability testing refers to is the ability to test for assurance of circuit lifetime and reliability at the wafer level. This includes the design of test and test structures that will test for known reliability failures on a wafer. Testing is usually done by using a Parametric Tester.

A parametric tester is a tester designed to test electrical parameters such as voltage, current, capacitance, etc. This type of tester differs from the testers used at the wafer sort and final test operations in the following way. Parametric testers do not test the functionality of circuits but test the electrical response of circuits. Parametric testers tend to do tests of a more analog nature. i.e. measurement of current, voltages, etc. whereas functional testers tend to be of a more digital nature. i.e. pass or fail.

To test on a parametric tester, the test structures (i.e., on the same wafer containing production die) and the tests are designed so that the test can be quickly done, usually in under a minute. The principle behind wafer reliability testing is nothing more than the continuation of building in quality, of being proactive rather than being reactive. With Wafer Level Reliability testing, rapid feedback is now possible. Wafer Level Reliability will not replace the traditional burn-in testing for infant mortality and long term life time testing, rather it will supplement this testing. It will provide additional data over a larger sample size that will ensure outgoing quality levels.

When there are major changes in the process, the effect on these changes have to be completely examined. Traditionally, this meant a life test through the use of accelerated burn in or stressing. The feedback could take months. With Wafer Level Reliability testing, rapid feedback is now possible. Wafer Level Reliability will not replace the traditional burn-in testing for infant mortality and long term life time testing, rather it will supplement this testing. It will provide additional data over a larger sample size that will ensure outgoing quality levels.

Throughout this dissertation, discussion has focused on what defects are and what is meant by reliability and yield defects. A yield defect is a killing defect that is detected either by the wafer sort testing process or the final package testing process. A reliability defect is a killing defect that causes the part to fail before the quoted lifetime, even though the part has passed all the testing methodology.

The question arises, what are these reliability defects and how can they pass the nominal testing methodology? Defects that cause reliability failures can be the same kind of defects that cause yield loss at wafer sort. They can also be different from the sort yield loss defects. How can the same defect be both a yield and reliability defect? To answer this, consider the example used previously of the killing defect caused by the metal bridging particle. Consider the same particle as discussed before, this time in the middle of a metal line, as shown in Figure 3.4.

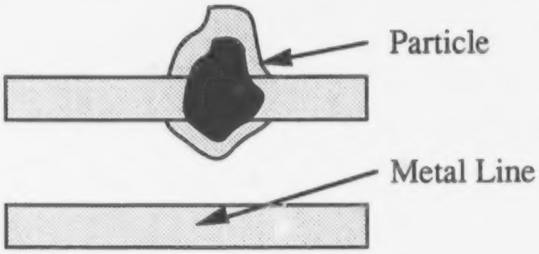


Figure 3.4. Potential reliability defect particle example

This particle is of the same type that causes metal bridging and shorts in the circuit, but its location would now be such that it causes no bridging between metal lines. The particle would then be positioned in such a manner that the amount of metal around the particle would be less than the normal amount of metal in a normal metal line. This particulate can then cause two kinds of reliability failures.

The process steps that caused the earlier examples of the metal particle defect are the same process steps that could have caused the defect in Figure 3.4. Even though this defect is not a defect that may cause failure at either wafer sort test or final test, it is still a concern because of the two reliability failures that can occur in the field. Some of the different reliability failures and how the manufacturing processes can cause them are mentioned below.

3.5.1 Electromigration and Stress Void Reliability Failures

In thin metal lines such as those used in semiconductor circuits today, there is a phenomenon known as electromigration. Electromigration refers to the movement of material within a structure such as a metal line due to the presence of an electric field. In thin metal lines, the higher the current density, the higher the electromigration. That is, the more that metal ions will physically move and migrate over time. The thinner the line, the more resistive it becomes. The more resistive a line becomes, the more joule heating occurs. The hotter a line is, the easier it is for ions to move. Therefore, electromigration is a catastrophic effect. The more metal migrates, the less metal there is to carry the current causing more localized heating causing more migration, causing higher current density, causing more heating, etc. The end result is that enough metal will migrate to cause an open in the circuit metal line(s) and the circuit will fail.

Electromigration may not necessarily be caused by a particulate defect. Any changes in metal composition will effect electromigration. Grain size is a factor in electromigration. As metal is deposited, metals such as Aluminium tend to grow from many sites at once. When the metal sites grow together, they form grain boundaries. Grain boundaries are where the crystal orientation is not the same, in the metal ionic structure. These boundaries are a result of the thin film deposition process used. Depending upon process conditions, the size of these grain boundaries can change. These boundaries can cause local stress and resistive points within a metal line.

But thin metal may not only be caused by a particle in the metal line. The line itself can be manufactured too thin and out of specification. The photolithographic process and the metal etch process can also cause thinning of the metal lines. In photolithography, if the focus of the stepper is not controlled, local thinning can occur. In Etching, if the etch profile is not controlled, thinning can also occur. Where the metal lines cross over topology is the part of the metal line most sensitive to this thinning. Again, having a basic knowledge of the semiconductor processing is critical to the identification of where the reliability defects occur and how to prevent them. Usually defects of this nature are detected by the constant testing and measurement of the critical process dimensions (such as metal line width) through out the process.

The equation which describes the effect on the lifetime of a metal line due to electromigration is given by: [5]

$$MTTF = A*J^{-n}e^{(Q/kT)} \tag{3.4}$$

where:

MTTF = Median Time to failure
A = determined by metallization
J = Current Density
Q = Activation Energy
k = Boltzman's constant
T = Temperature (°K)
n = Current density factor

This equation shows that the lifetime of a metal line is affected by the temperature of the line and the current in the line.

Another possible failure that can occur in the example of the particle in a thin metal line is stress voiding. Different materials in semiconductor manufacturing have different stresses. Metals tend to have compressive stresses whereas insulators such as silicon dioxide tend to have tensile stresses. When you have stresses of these opposing natures, there can be failures over time due to these stresses relieving themselves with the circuit.The metal lines tend to pull apart and fractures open up in the lines. The stress voids are formed almost immediately, within 24 hours after completion of processing. If the voids are large enough, the circuit will fail immediately. If they are not, the metal line will be thinned. This thin portion of the metal line can experience an electromigration early lifetime failure.The way to prevent this type of failure is to control the stresses of the films during the deposition steps.

The effect of stress voids on a metal line is illustrated in figure 3.5.

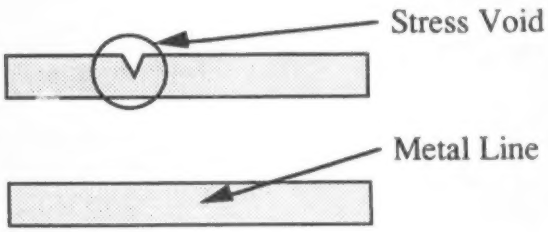


Figure 3.5 Stress void in a metal line

With the metal particle example of figure 3.4, stresses can occur in the metal around the particle. The important thing to observe is that the particle of our examples is a type of defect that can cause many different type of failures, depending upon where it occurs and other factors present in the processing of the circuit. The failures that occur at production testing are different than the failure mechanisms that cause early life time wear out. The important thing is that this one type of defect can cause many different types of circuit failure. Stress voiding can occur where ever the metal thins in the circuit. If the circuitry under the metal has a lot of topology, that is, not planar, the circuit will be more sensitive to stress voiding than a circuit that has a very planar surface under the metal. If a thin metal line is caused by either etching or photolithographic problems, that line will be more sensitive to stress voiding and subsequent electromigration failure over time.

3.5.2 Transistor and Dielectric Reliability Failures

There are additional failures that can occur after wafer sort and final testing. Failures can occur in the transistors or in the dielectric thin films that separate the conductors. The reliability failures in transistors can occur when there are improper doping of the P or N type semiconductors or impurities are trapped in these regions. The level of doping or impurities may be such that the IC will pass the wafer sort test and the final test, but the electrical characteristics of the transistors will be marginal. During repeated switching and stressing that occurs during normal operation, there may be migration of the dopants such that failure will occur. The failure will be of the early wear out variety.

Another similar example is if there are contaminations in the thin oxides of MOS circuitry, especially in the gate oxide. This is an example of dielectric reliability failure. The "capacitor" portion of the gate to substrate region has to charge or discharge correctly and not leak current. If the plasma etching processes or the plasma deposition processes (PECVD) damage this sensitive oxide, failure can occur. Because etchers such as RIE etchers have a high bias, (as discussed previously) ions and electrons can penetrate into the gate oxide. The etcher can act as an undesirable ion implanter, doping the gate oxide with impurities. It is these impurities that can change the dielectric properties of the gate oxide and therefore the insulator properties. This change affects how the transistors switch. The CMOS invertor example may not have one transistor switch off at all. The IC will then draw too much current and heat up. The IC will probably fail functionally.

If the impurities are of a high enough dose, the die on the wafer will draw too much current and fail the power up test at either wafer sort testing or final testing. If the impurities are of such a level that the circuit passes these tests, early wear out failures may still occur in the field. This increased gate leakage over time can occur in the transistors due to repeated stressing of the oxides. The oxides can be stressed due to repeated switching of the transistors. The level of contamination in the gate oxide may be of a level that is too low to cause failure at wafer sort or final test, but the transistor may degrade over time such that the part will fail earlier than the published lifetime.

3.6 Reliability Defect Detection and Testing

The defects that tend to be found in reliability failures can either be of the global or of the point type. Some reliability defects are very borderline yield defects. The defects are such that they barely pass the wafer sort and final test on the testers. These defects generally will fall out in the initial infant mortality burn ins. With the increase in tester sensitivity and the decrease in the defect densities, more failures are being caught at the tester and fewer failures are being caught at infant mortality burn in. The failures that do get through tend to be more of the electromigration, gate oxide integrity and stress voiding type that do not show up at infant mortality. Circuits today have smaller dimensions and lower defect densities than ever before. These ICs are more sensitive to early lifetime wear out. Hence, the focus with the semiconductor industry is now shifting towards these life time limiting failures.

A table of common defects and their effects on circuits is given. Table 3.1 is by no means complete and gives an illustration of the types of defects in semiconductor processing. Each factory and associated processes will have their own unique failure mechanisms and associated defects.

Table 3.1 Common yield and reliability defects

Defect Type	Yield Affects	Reliability Affects
Contamination, Particles	Point Defects	Point Defects
Misalignment (from Steppers)	Large Misalignment causes Point Defects sometimes Global	Small Misalignment can cause Point Defects
Misprocessing	Yes	Yes
Doping problems	Yes, usually Global	Yes, usually Global
ESD damage	Yes	Yes
Radiation Damage (Plasma Processing)	High damage	Oxide degradation

As was illustrated previously, the defects that cause yield problems can cause reliability problems. Usually the defects that show up in the sort yield are of a more severe nature than the defects that cause reliability failures, weather or not they are infant mortality or early wear out defects. The goal today is to produce good yielding and reliable parts. If a company does not do this, they will be out of business. The semiconductor industry is very competitive. The detection and prevention of defects that cause reliability failures is an on-going task. The best way to prevent defects is to minimize the source of them in the manufacturing process. This is done by being proactive rather than reactive. Quality is built into the process through wafer level reliability, in line process monitoring and control and traditional reliability testing.

Table 3.2 lists tests that have been developed to detect reliability failures [10].

Table 3.2 Wafer reliability tests

Wafer Level Reliability Tests	**Failure Type**
Electromigration SWEAT	Metallization / Electromigration
Electromigration BEM	Metallization / Electromigration
Contact Electromigration	Metallization / Electomigration
Contact Spiking	Metallization / Electromigration
Via Voiding	Metallization
Metal Stress Migration	Metallization / Stress Voiding
Ramped Voltage QBD	Dielectric
Interlevel Dielectric	Dielectric
Top Passiviation	Dielectric
Oxide Charge Trapping	Dielectric
Interface Trapped Charge	Dielectric
Secondary Slow Trapping	Transistor
In Process Charging	Transistor / Dielectric
Mobile Ion Contamination	Transistor
Hot Carrier	Transistor

The wafer level reliability tests illustrated in Table 3.2 are very quick and typically inexpensive when compared to the traditional methodologies of testing semiconductor reliability failures. They are designed to test for reliability failure mechanisms. Results of the tests are known in seconds as compared with months for the more traditional burn in techniques.

Again, these tests supplement the more traditional burn in and life time techniques. The more traditional reliability tests are needed to identify the actual reliability failure modes. From this data, wafer level reliability tests can be designed and implemented to provide a higher level of screening for specific reliability failures. This higher level of sampling will guarantee a higher level of out going quality of the ICs.

3.7 Summary

In semiconductor ASIC manufacturing, the yield and reliability of the parts have been traditionally treated as separate issues. It was believed that the defects that caused yield problems could be detected at the wafer sort step and the reliability of the parts was tested through burn in procedures to control infant mortality and collect data on circuit lifetime. The feedback was then to the engineers involved with the processing set up and control. The processing engineers tended to be more concerned with the immediate yield concerns and keeping their processes in control.

Traditionally reliability tasks involved initial burn in of finished parts to detect and weed out infant mortality failures and involved life time testing though accelerated temperature shock. The wafers tended to be sampled from lots and the life time tests results were not known for a period of months.

But it has been shown that the industry has changed. Some of the assumptions that were valid five and ten years ago are now questionable. The level of defects was such that most of the failures tended to be infant mortality. The parts are now more sensitive due to the smaller geometries. The defect densities have decreased to enable semiconductor factories to get yields on devices that were considered impossible then. This has meant that there has been a shift towards more early wear out reliability failures being more dominant than the infant mortality failures that was seen back then. Because the dimensions are smaller, the parts are more susceptible to failures associated with current densities, materials, material stresses, etc.

Another issue here is that more processing is now done in a single wafer fashion than ever before. The sampling plans for reliability were generally based on a lot by lot bases. Now the processes that individual wafers may see within a lot may have more variability than the variability from lot to lot. Some sampling plans that are based on lot variation may now be invalid.

One of the biggest factors influencing the reliability picture is that customers are demanding higher quality than ever and higher assurances of quality at an ever cheaper and cheaper cost. To insure higher levels of quality through testing and yet continue to reduce the cost of the final product is an almost impossible task. To sample more for the traditional way of reliability testing costs more.

As the technology complexity increases and device geometries shrink, new processes and new materials will be used in the construction of semiconductor devices. With these new frontiers will come new failure mechanisms and types of defects. The on-going challenge will be to identify these defects and minimize their impacts on yield and reliability through changes in design, procedures and processes. To identify the defects, a basic knowledge of the semiconductor processes is required. Understanding of the photolithography, the ion implantation and diffusion, the etching and the thin film deposition is critical in the identification and prevention of yield and reliability defects. The same defect can be caused by different processes in the manufacturing.

Knowledge of the basic structure and operation of the transistors is critical to the understanding of defects. Transistor operation gives insight to operation of the circuit. The tester will determine how the part is failing. Knowledge of the transistor operation may assist in identifying the actual defect that causes this failure mode.

Testing of the circuitry adds complexity to the identification and prevention of yield and reliability losses. The ability of the tester and the test program to properly identify bad Integrated Circuits must be constantly questioned, especially with today's complex circuits.

The on-going identification of killing defects and non-killing defects, whether they be of the point or global defect type, is critical. The non-killing defects at wafer sort and final test may fail in the field causing reliability early wear out failures. Different techniques such as Yield Modeling have been shown to be important in the identification and comparison of defects between Integrated Circuits of different sizes. Models such as the Murphy Seeds and the Poisson yield models have been shown to be good tools when comparing actual yield data.This comparison is critical in identifying specific circuits that may have yield and reliability problems.

As the complexity of the parts increase, so does the complexity and the amount of testing required to insure that the high levels of quality are met. The challenge is to do this and yet have a cost competitive part. Quality testing is important at all levels of semiconductor processing. Testing the process through in line monitors and tests is a vital part of SPC (Statistical Process Control) and important to the prevention of yield and reliability defects. Proper testing of the finished wafers at wafer sort testing and the chips at final test is critical to the prevention of defective parts ending up in final electronic products. Analysis of the failures at these tests help to identify and prevent further yield loss defects.

A proper reliability program involving traditional reliability testing for infant mortality failures and lifetime testing along with a rigorous wafer level reliability testing program will provide the testing that will insure product lifetime and quality. All the issues discussed previously are even more important when Application Specific Integrated Circuits (ASICs) are considered. Each individual design is specific to a customer. An ASIC manufacturing facility will have hundreds and even thousands of designs. This puts more demands on the previously mentioned testing programs.

It is through the application of SPC principles in manufacturing, yield monitoring and failure analysis, reliability testing through traditional testing and wafer level reliability that the quality of ASIC Integrated Circuits will be realized insuring competitiveness in the changing marketplace. The yield and reliability of the finished parts will meet the tough performance, cost and quality demands that customers of ASIC VLSI ICs are demanding today.

CHAPTER IV

ELECTROMIGRATION FUNDAMENTALS

4.1 Introduction

Electromigration has become one of the most critical failure phenomenon seen in the failures of today's complex semiconductor circuits. Electromigration is a physical effect exactly like the name implies - the physical migration of ions due to the presence of an electric field. During the course of the "lifetime" of an integrated circuit, current flows through microscopic circuit interconnections. This current sets up an electric field that "drags" along the ions that make up the metal interconnection. Over a period of time, enough ions will be transported to cause a failure by opening an area within the line. This failure is an electromigration failure, a catastrophic effect.

To obtain today's desired performance, the size and geometries of the transistors and their associated interconnections have decreased. These smaller geometries have brought new problems and challenges to the forefront. With larger geometries, the lifetimes of metal interconnections was never a problem. But with the smaller geometries, the lifetimes of the metal interconnections have become a factor. Electromigration is more sensitive to smaller geometries. Therefore, as the line widths have shrunk, the effects on circuit lifetime due to electromigration has increased.

Some electromigration tests are done today at the wafer level. By the use of accelerated lifetime testing at the wafer level, manufacturing processes are monitored for electromigration failures. But the wafer level tests do not fully replace the more traditional life time stress tests. Semiconductor manufacturers still sample and stress finished product for reliability failures. The failures from these lifetime stresses (burn in) are then analyzed to see if the failure mode is due to electromigration failures or some other failure modes. A complete understanding of the actual physics of the electromigration phenomenon is not fully known at this time. The present theories, test methodologies and assumptions will be discussed in this chapter. The integration of electromigration lifetime testing into semiconductor manufacturing processes will be reviewed.

4.2 Electromigration Physics

To properly understand the testing and reliability of electromigration tests, an understanding of the physics of the electromigration phenomenon is required. In semiconductor manufacturing, the interconnection of transistors is done through conductors that are formed into lines through the manufacturing process. The conductor is usually first deposited onto a wafer. The wafer then has the pattern for the interconnection transferred onto the wafer by photographic and etching processes. This removes the unwanted conductor material, leaving the interconnecting conducting line.

The electromigration effect is dependent upon the structure of the conducting material. To understand the effect, a basic understanding of the deposition and structure of the conductor is required. In silicon semiconductor manufacturing, most interconnecting conductor lines involve aluminium metal or polysilicon material as the lines or a major component of the lines.

4.2.1 Structure of Conductors

Most deposition of thin film metals or conductors of the type that are used in semiconductor manufacturing is done under high vacuum conditions. The solid metal is deposited or grown onto silicon wafers, as reviewed in Chapter II of this thesis. During Deposition, metals tend to orientate themselves in ordered arrangements in a crystal lattice structure. This lattice structure typically consists of three types, typically - polycrystalline, amorphous, and single crystalline. Single crystal layers are grown from a seed crystal and the orientation within the lattice is that of the single seed crystal. Polycrystalline structures have many different orientations within the thin film and are formed by individual crystals being deposited and growing larger until all the crystals come together to form a single film. Amorphous structures are thin film layers that do not have a crystalline structure or orientation of the individual atoms.

Most deposition in semiconductor IC manufacturing is of the polycrystalline type. The metal starts to form or grow from many individual islands. The orientation of the crystal structures in each of the islands is not the same. As the "islands" come together, there is a discontinuity in the crystal lattice structure. This is illustrated in Figure 4.1 below.

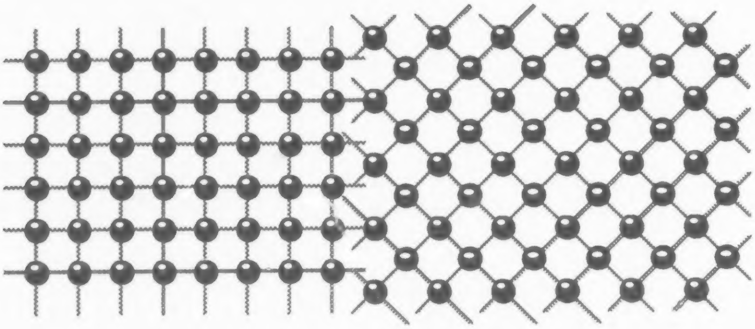


Figure 4.1 Crystal lattice structure

This Figure shows the structure within a conductor. There is a discontinuity between the two orientations of the atoms between the two “islands”, as illustrated above. This discontinuity in the structure is referred to as a grain boundary. The orientation of the solid metal crystal or "grain" is different in the two regions of figure 4.1. The size of the grains are dependent upon various process conditions. The size of the grain boundaries are approximately five to ten Angstroms [11] and grain size is typically from a few hundred to a few thousand Angstroms.

This grain boundary is very important in the study of electromigration. At this location the atoms and ions are not bound by neighbouring atoms or ions. This means that the forces that bind the atoms/ions are less and therefore, it takes less force for the atom/ion to migrate or diffuse through the lattice structure. Therefore, to understand electromigration, an understanding of the bonding forces that prevent the migration of ions is important.

4.2.2 Ion Potential Bonding and Diffusion

Within a metal crystal, the ions are bound together by the forces that result from the ions sharing their valence electrons with the entire aggregate [12]. The binding force that holds the atoms together is opposed by the force exerted by the mutual repulsion of the ions and their associated completed electron shells. The atoms in the crystal can, therefore, be considered to be in a potential well within the three dimensional crystal array or lattice. Each atom can vibrate thermally within this potential well but can only move a short distance.

But atoms within a crystal lattice do not only exhibit potential energy, but also kinetic energy. The vibration of the ions within the crystal lattice is a form of kinetic energy. The temperature of the material in which the atoms are located is determined by the kinetic energy of the thermal agitation of the atoms. As the atoms vibrate in their potential wells, the energy changes from purely potential energy to fully kinetic energy [12]. At the bottom of the potential well - the location in the crystal lattice where the atom stops movement - the potential energy is equal to kT where k is Boltzsman constant and T is the absolute temperature [12]. As the atom moves up the potential well, the energy becomes kinetic, until when it is at the top of the well, the energy is all kinetic and also equal to kT [12]. With the atoms vibrating, some are in the fully potential energy state and some are in the fully kinetic, with the rest of the atoms somewhere in between. Therefore, the temperature of the material is the mean value of the kinetic energy, which is 1/2kT [12]. The distribution of these energies are distributed according to the Boltzmann distribution.

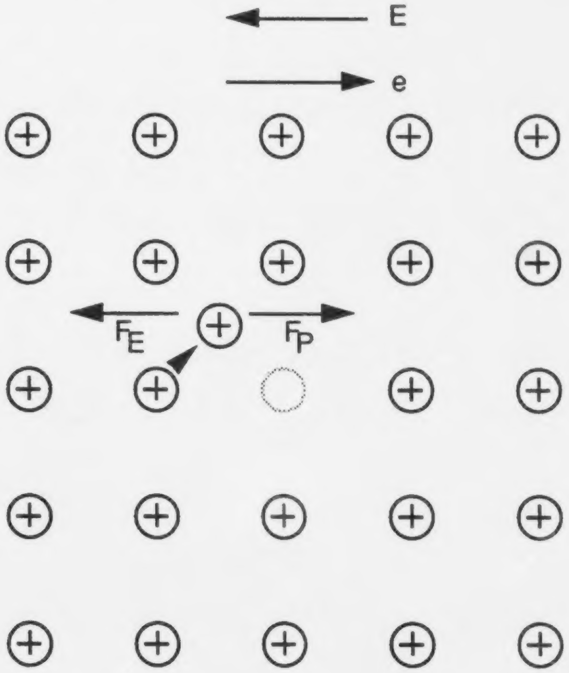
According to the Boltzmann distribution, the proportion of atoms, p, of a material that have a kinetic value greater than Q is given by [12]:

$$p = e^{\left(\frac{-Q}{kT}\right)} \qquad (4.1)$$

This equation shows that the number of atoms that have an energy greater than a particular value is determined by the temperature of the material in question.

The understanding of the energy is important to show the mechanisms of diffusion of the ions. When the ions are at the top of the potential energy well, they may possess enough energy to escape the potential energy well or the forces that bind them into the crystal lattice. At any temperature other than absolute zero there is always a percentage of metal ions within the crystal lattice that possess this sufficient energy [12]. This effect is dependent upon temperature, as illustrated by the above equation.

The ions or atoms that are in this energy state are free of the crystalline lattice structure and are "activated". These activated ions can then undergo a random rearrangement within the crystal lattice. This process of diffusion which takes place under no concentration gradient or chemical potential is referred to as "self diffusion" This is illustrated in Figure 4.2.



An ion in the process of exchanging position
with a near neighbor vacancy.

Figure 4.2 Ion self diffusion diagram [12]

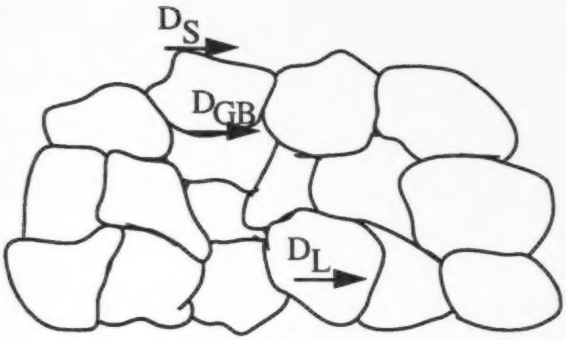
For diffusion to take place, not only must an ion be able to have enough energy to overcome the forces that bind it into place, there must also be a vacancy in the ion structure for the ion to migrate to. In close spaced crystalline structures, the ion exchange takes place with a near neighbour. This means that not only must an ion be "activated", there must be a near neighbour vacancy for the ion to move into the vacancy. This means that the ion exchange involves two steps and associated energies. The activation energy for vacancy formation, Ef, is the energy required for a near neighbour ion to be missing or vacant. The higher the temperature of the material, the more sites are vacant. The other energy, Ej, is the energy required for an ion to jump from its location or potential energy well into a vacant neighbour location or well. The sum of these energies gives the total energy required for self diffusion. [12]

$$\text{Qdiff} = \text{Ef} + \text{Ej} \qquad \textbf{(4.2)}$$

Aluminium is an example of a closely spaced ion structure in which this self diffusion phenomenon occurs. Through experimental studies [12], it has been shown that for Aluminium, Ef=0.73 eV and Ej=0.75 eV. The activation energy for self diffusion of Aluminium ions through the crystal lattice is therefore 1.48 eV. The diffusion or jumps takes place in a random fashion in the crystal lattice and is independent of previous jumps. In aluminium, this means that a ion can jump to any of the 12 possible near neighbour vacancies.

In polycrystalline films, there are three kinds of diffusion that can take place. The diffusion can take place within the crystalline lattice structure, called lattice diffusion (D_L). Diffusion can take place along the surface of the material, called surface diffusion (D_S) or Diffusion can take place along grain boundaries, called grain boundary diffusion (D_{GB}). At the grain boundaries, there is a disruption in the crystal lattice structure. This means that there are fewer near neighbour ions than within the crystal lattice. With fewer near neighbour ions, there is less force holding ions at grain boundaries in place and, therefore, the potential wells are shallower.

Experimental measurements have determined that the activation energy for diffusion of aluminium along a grain boundary to be 0.55 eV, which is less than the self diffusion energy required within the crystal lattice. The grain boundaries provide an easy path for self diffusion. The surface of the metal also has fewer near neighbour ions and the diffusion is affected the same way as across the grain boundary. An illustration of the three types of diffusion within a crystal lattice of aluminium is shown in Figure 4.3. [12]



ALUMINIUM CRYSTALLITES

D_S SURFACE DIFFUSION
D_{GB} GRAIN BOUNDRY DIFFUSION
D_L LATTICE DIFFUSION

Figure 4.3 Types of Diffusion within a Thin Film

4.2.3 Ion Electromigration

There is a diffusion of ions within a crystal lattice. This diffusion is random and there is no net transport of ions. But if the material in question happens to be a conductor with electrons flowing through it, the phenomenon called electromigration occurs. With the application of current, the self diffusion effect changes into a directional diffusion with the presence of an electric field and the flow of charge carriers. This is shown in Figure 4.2 The electron flow and the direction of the Electric Field is illustrated. Since the diffusion of ions is now directional, there is a net diffusion in a single direction and a net migration of ions or mass. Because this migration phenomenon is caused by the influence of an electric field, the phenomenon is referred to as electromigration.

There are two forces that cause electromigration. The first force is due to the positive ion interaction with the electric field. (FE). The second force arises from the charge carriers giving up momentum to the ion (FP). The charge carriers can either be electrons for n type conductors and holes for p type conductors. The force, FP is in the direction of the charge carriers. The electric field force is in the direction of the field.

The electric field is in the opposite direction of the electron flow. Therefore, the net force on an n type conductor (electron conductor flow) is then [12]:

$$F = FE - FP \tag{4.3}$$

The accepted theory is that the shielding electrons causes the force due to the electric field to be small (FE) in comparison to the force due to the carrier momentum loss (FP). In good electrical conductors such as metals, FP is the dominant force [11]. Therefore, the assumption is that the force on the ion is essentially:

$$F = FP \tag{4.4}$$

As the carriers travel through the conductor, momentum is loss to the ions. The flow of the carriers is referred to as the "electron wind". FP has been referred to as the frictional force due to the electron wind.

From basic physics, the force on a charged particle in an electric field is written as:

$$F = qE \tag{4.5}$$

Now, if Z*e is the effective charge assigned to the migrating ion, then the net force can be written as:

$$F = Z^*eE \tag{4.6}$$

The effective charge assigned to the migrating ion, Z*, is given by:

$$Z_l^* = z\left[1 - \gamma\left(\frac{\rho_d}{N_d}\right)\left(\frac{N_l}{\rho_l}\right)\frac{|m^*|}{m^*}\right] \tag{4.7}$$

where:

l refers to the lattice parameters
z refers to the number of conduction electrons per atom

$(\frac{\rho_d}{N_d})(\frac{N_l}{\rho_l})$	refers to the ratio of the specific resistivity of the moving defects (migrating ion) to the resistivity of the lattice atoms {d for defects, l for lattice}
$\frac{\lvert m^*\rvert}{m^*}$	this term is used to determine the direction of the force due to the sign of the charge carriers
γ	is an averaging term which accounts for the variation of force along the length of an elemental atomic jump, or distance between atoms in a crystal lattice. This is usually assumed to be 0.5 [11]

Up until now the development has been along the force on a migrating ion. Practically, the force on a migrating ion is not directly measurable. But by using the Nernst-Einstein equation, the drift velocity can be related to the force by [12]:

$$\vec{v} = MF \tag{4.8}$$

where:

$\vec{v}$ = ion drift velocity
M = ion mobility
F = force

Now, the mobility M is given by [12]:

$$M = \frac{D}{kfT} \tag{4.9}$$

where:

D = self diffusion coefficient
k = Boltzmann's constant
f = A correlation factor depending upon the lattice type

Now the self diffusion coefficient has a temperature dependence and is often expressed as [2]:

$$D = D_0 e^{-(Q/kT)} \tag{4.10}$$

where:

D_0 = the frequency factor (in cm2/s)
Q = Activation Energy (in eV)
k = Boltzmanns Constant
T = Temperature (in K)

The mobility of an ion can not be directly measured. However, the ion flux can be observed. And the ion flux can be observed through indirect observation of the effects of the ion flux such as void formation and hillock growth. These formations occur at regions of divergence of the ion flux. This divergence is the change of electron flow across a region such as a grain boundary. Because of the changing lattice structure at a grain boundary, the electron flux will diverge at this change in the lattice structure.

The ion flux can be expressed as [12]:

$$J_{ion} = N\vec{v} \tag{4.11}$$

Now, substituting in the mobility expression from equation 15 yields:

$$J_{ion} = N\frac{D_0 e^{-(Q/kT)}}{kfT}F \tag{4.12}$$

And substituting the force from equation 4.6 yields:

$$J_{ion} = \frac{ND_0 e^{-(Q/kT)}}{kfT}Z^*qE \tag{4.13}$$

If the electric field in the above equation is substituted by:

$$E=\rho J \tag{4.14}$$

where

ρ = volume resistivity
J = current density

which results in:

$$J_{ion} = \frac{ND_0 e^{-(Q/kT)}}{kfT}Z^*q\rho J \tag{4.15}$$

This predicts the ion flux as a function of the current flux to the first power. This result corresponds to observations made in bulk metal [11].

4.3 Electromigration MTTF

Electromigration affects the lifetime of metal conductors in semiconductor circuits. Therefore, the important relationship is the Median Time to Failure (MTTF), or the lifetime of 50% of a population. The work done by James R. Black has shown that the MTTF is proportional to the cross sectional area of a film and has shown that as a first approximation, it can be assumed to be inversely proportional to the divergence of the ion flux as shown below [12]:

$$\text{MTTF} \ \alpha \ \frac{A}{\nabla \cdot J_{ion}} \ \{\nabla \cdot J_{ion} = \text{DIFF} \ J_{ion}\} \tag{4.16}$$

Now, the ion flux equation 4.10, has the assumption that there were no temperature gradients. Where the divergence of the ion flux, where the vacancies are trapped and where the ion flux is due to structural variations, the divergence of the ion flux is given by [12]:

$$\nabla \cdot J_{ion} \ \alpha \ \frac{J}{T} D_0 e^{-(Q/kT)} \tag{4.17}$$

Therefore, if the assumption above is that there are no temperature gradients and the divergence of the flux is only due to the structural variations within a conducting stripe, the MTTF becomes [12]:

$$\text{MTTF} = \frac{KAT}{J \ D_0 e^{-(Q/kT)}} \tag{4.18}$$

But the assumption that the temperature gradient does not affect the electromigration does not hold true. Joule heating effects within a thin metal line was shown to produce a divergence in the ion flux which is proportional to the current density cubed as shown below [12]:

$$\nabla \cdot J_{ion} \ \alpha \ \frac{J^3}{T^3} D_0 e^{-(Q/kT)} \tag{4.19}$$

and substituting in to the MTTF formula yields [12]:

$$\mathrm{MTTF} = \frac{K^*AT}{J^3 D_0 e^{-(Q/kT)}} \qquad \textbf{(4.20)}$$

Now since the region of temperature in which the metal line does not melt and which electromigration occurs is fairly close to room temperature, the top temperature term T is usually included in the constants K, K* and the Area A , the diffusion frequency factor Do and is given by a single constant A*. Combining both equations and rearranging gives [12]:

$$\mathrm{MTTF} = A^* J^{-n} e^{(Q/kT)} \qquad \textbf{(4.21)}$$

Now this equation has been considered to be the industry norm and is quoted in most literature that deals with electromigration phenomenon. This shows that the MTTF of a metal conductor is proportional to the current density in the metal conductor J to the power of -n and to the diffusion of the ion as a function of temperature. The power of J is usually given to be between -1 and -3.

Experimental results quoted by James R. Black, show that the power of J in the previous equation to be either -2 or -3. This was observed for thin films with a current density in the 10^5 or 10^6 A/cm^2 . It is in this region of current densities that there is enough energy to heat and stress the line such that electromigration can be observed with the proper heat sinking on the sample, but not enough such that the Joule heating could melt the metal line . In samples of bulk metal at low current densities, (10^3 to 10^5 A/cm^2), the power of J in the equation was found to be -1.

At the lower current densities, the accepted theory is that the majority of migration is due to, firstly, ions having to overcome the atomic forces that bind them into the crystalline lattice and secondly, migrate to a near neighbour site. As has been illustrated before, this requires an input of energy. The only divergence of the ion flux in this case is due to the structural variations in the metal line. At the higher current densities, some ions now have enough energy to be free of the lattice structure and be unbound. They will migrate easier for they only require enough energy to move to a vacant site in the lattice structure. Here there will be temperature gradients within the line itself. The electromigration effect occurs where these gradients in structure and temperature occur.

In practice, within any given metal lines there are ions within the lattice structure that have enough energy to be unbound and some that do not. At the higher temperatures and current densities, there are more unbound ions. With a higher current density, there will be more electrons imparting more force on the ions in the lattice structure, transferring energy and momentum to the ions in the crystal lattice. This is why at the higher current density, the power of J will appear to get closer to -3 as the temperature and/or the current density increase.

4.4 Electromigration Test Techniques

Electromigration is only one part of a series of long term and short term tests that are used to monitor and determine reliability of semiconductor parts. If the parts contain defects such that the functionality of the circuit is in question, then the part will fail during the initial testing done at the end of the manufacturing process, during either the wafer sort testing or the final testing. However, some reliability defects are very borderline yield defects. The defects are such that they barely pass the wafer sort and final test on the testers. These defects generally will fall out in the initial infant mortality burn ins. With the increase in tester sensitivity and the decrease in the defect densities, more failures are being caught at the tester and fewer failures are being caught at infant mortality burn in.

The failures that do get through the functional testing tend to be more of the electromigration, gate oxide integrity and stress voiding type that do not show up at infant mortality. Circuits today have smaller dimensions and lower defect densities than ever before. These ICs are more sensitive to early lifetime wear out. Hence, the focus is now shifting towards these life time limiting failures.

Electromigration is only one of the tests that are done to measure lifetime of semiconductor products. Electromigration testing at the wafer level was first done by stressing parts or test structures through the application of stress through the use of a constant current. The testing apparatus in this case consists of a standard parametric tester that includes power supplies and digital multi-meters within the tester connected to a heated chuck wafer prober. The wafer is heated to a desired temperature and a known current density is applied. The current densities in this test are in the range of 10^3 to 10^5 A/cm^2. This is enough energy to heat and stress the line such that electromigration can be observed, but not enough such that the Joule heating becomes a significant effect. The wafer is heated to apply additional stress. Even with this acceleration, the test takes between 100 to 300 hours, before electromigration failure is observed. Therefore, this type of test takes to long to be an effective process line monitor. However, the results have been shown to give fairly good MTTF estimates.

4.4.1 SWEAT Test

One of the two most widely used acceleration techniques for determination and testing for electromigration is the S.W.E.A.T. test. **SWEAT** stands for **S**tandard **W**afer level **E**lectromigration **A**cceleration **T**est. This test technique was developed by Bryan J. Root and Tim Turner when they were at Mostek Corporation. The reason that this particular test was developed, was to allow electromigration testing to be done with the use of parametric testers with the total test time to be under two minutes. Before this method was invented, electromigration testing was done by heating a wafer and applying a constant current, as described previously. Even though the effect was accelerated, by the hot chuck method, the MTTF of the line was still in the neighbourhood of a few days. To make the test more practical in a manufacturing environment, a higher current density was required.

The draw back of this type of a more accelerated testing is that with a higher current density, above 10^6 A/cm^2, Joule heating effects become significant. This can lead to inaccuracies in the measurement of the MTTF, as given by the "accepted" formula seen previously:

$$MTTF = A^{*}J^{-n}e^{(Q/kT)} \qquad \textbf{(4.22)}$$

To produce electromigration failures in the time required for production line testing, the hot chuck method can not be used. The temperatures that the wafer would be heated to would be close to the melting point of the line. Also, heating the entire wafer could possibly damage the wafer with the stresses caused on the wafer by the mismatch of the thermal expansion coefficients of the materials involved in the physical makeup of the circuits. This is because different materials expand and contract at different rates as the temperature either rises or falls. The heat will age any production circuits next to the test structure by stressing the circuits thermally. By heating and cooling of the wafer for electromigration tests, there could be failures caused by the stresses on the lines themselves.

The SWEAT technique uses controlled joule heating on a metal line in a test structure. This is difficult to do. Joule heating gives rapid temperature rise in a line, but this temperature is extremely difficult and effectively impossible to monitor. The resistance of a metal line is modelled to be [13]:

$$R = R_0 + \Delta R_A + \Delta R_L + \Delta R_T + \Delta R_{EM} \qquad (4.23)$$

where:

R_0 = Initial resistance of the line

ΔR_A = Change in resistance of the line due to Annealing of the metal

ΔR_L = Change in resistance of the line due to Joule heating

ΔR_T = Change in resistance of the line due to external applied heat

ΔR_{EM} = Change in resistance of the line due to electromigration

This formula means that it is impossible to use the resistance of the metal line under test as the method of determining the instantaneous temperature of the line.

The SWEAT test methodology uses power density as the line monitoring factor. This is done by first determining the metal line temperature as a function of the power density. Work done by Turner and Root on the development of the SWEAT structure shows the power density as a function of line temperature. The approach used to determine the relationship between power density and temperature was to first measure the change in resistance at different temperatures. A typical line in the SWEAT test structure was measured at different temperatures. The line was constructed on a wafer. The temperature of the line was set by raising the temperature of the chuck that the wafer sat on.

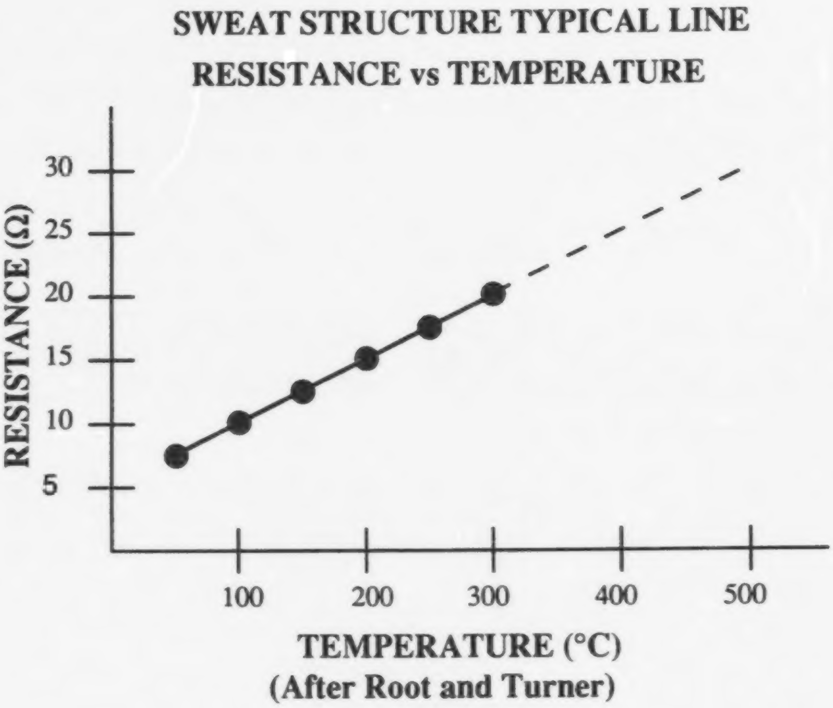


Figure 4.4 SWEAT structure resistance/temperature relationship

The above graph is typical of the results of this test. From this resistance and temperature relation, the following equation is given [13]:

$$R = R_0 + \Delta R = R_0[1 + \beta(T - T_0)] \qquad \textbf{(4.24)}$$

where [13]:

$$\Delta R = R_0 \beta \Delta T \qquad \textbf{(4.25)}$$

and:

R_0 = Initial line resistance
R = Resistance of the line at temperature T
β = Thermal coefficient of resistance
T = Temperature of the line
T_0 = Initial temperature of the line

From the graph above, this is shown to be a linear relationship. From this relationship, it is then possible to determine the temperature of the line due to the resistance of the line. By monitoring the current through the test line and the voltage across the test line, the resistance of the line and, therefore, the temperature of the line is readily available.

Turner and Root then sent a power pulse through the test structure for approximately 100ms. From the power applied and the change in resistance the following graph was obtained:

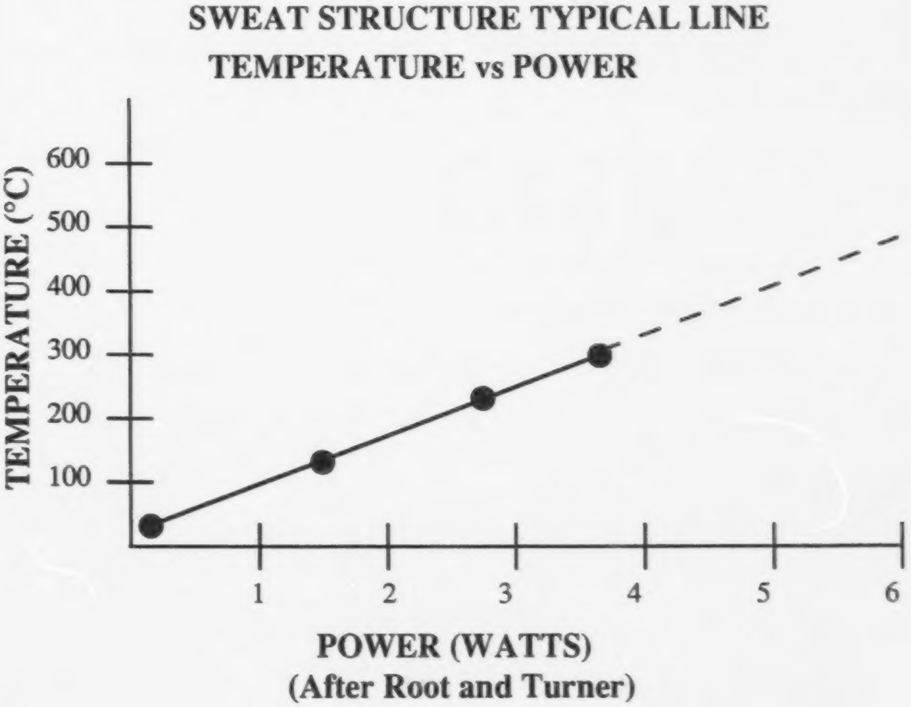


Figure 4.5 SWEAT power/temperature relationship

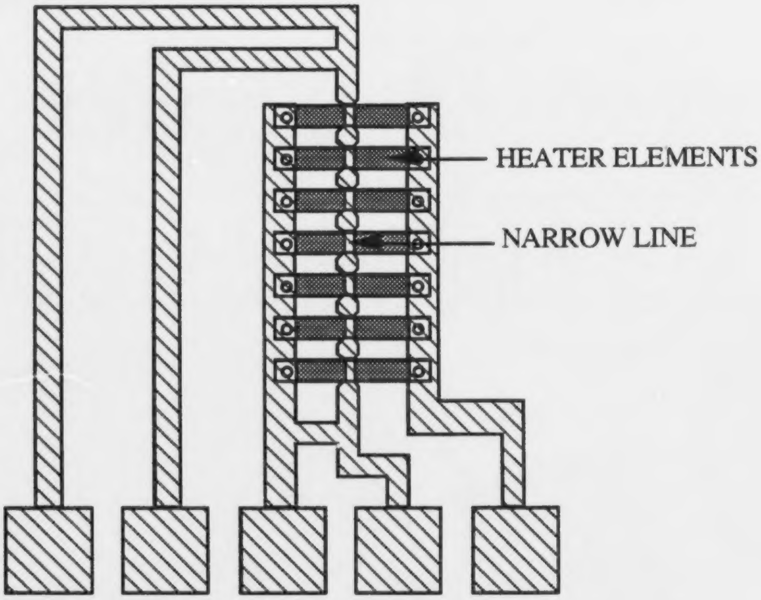
In all of their experiments, the relationship between temperature and power was found to be linear. This means that the relationship between power and temperature can be expressed as [13]:

$$T = P_D Q + T_D \qquad \textbf{(4.26)}$$

where:

P_D = Power Density
Q = slope of the line given in the graph above

The test structure used in the SWEAT tests consists of both narrow and wide lines. A typical layout is shown in Figure 4.6.



SWEAT TEST STRUCTURE
(After Root and Turner)

Figure 4.6 Typical SWEAT test structure [13]

The SWEAT structure consists of both narrow and wide lines. The reason that two different line widths were chosen is that this type of structure is useful in showing electromigration failures. Electromigration will occur where there is an ion flux gradient or where there is a temperature gradient. The SWEAT structure creates these gradients by having both the "large" and the "narrow" metal widths. The current density is much higher in the narrow lines. The "large" heat sink thick structures are cooler than the narrow lines. This is the necessary gradient to precipitate electromigration failures. Therefore, this circuit attempts to create the worst case scenario for electromigration structures within a circuit. Also, when the narrow lines cover an underlying structure, the narrow lines to be higher than the rest of the line or "step" over the underlying structure. This "step coverage" is more sensitive to changes in the photolithographic and etching processes. This narrow line that is on top of the underlying structures or the "heaters" will, therefore be, more sensitive to slight process shifts more than other elements in the test structure. Therefore, the "narrow" lines are the most sensitive location for void formation due to electromigration effects. This line stepping over underlying structures is a common occurrence of failure in actual circuits.

The work done by Turner and Root as demonstrated by the previous graphs has shown that the power density is proportional to the line temperature. Therefore, the relationship between the change in temperature of the wide lines and the narrow lines is given by:

$$\Delta T_W = (N_{SW}/N_{SN})\Delta T_N \qquad \textbf{(4.27)}$$

where:

ΔT_W = Change in temperature of a wide line
N_{SW} = Number of squares in a wide line
N_{SN} = Number of squares in a narrow line
ΔT_N = Change in temperature of a narrow line

Since $N_{SW} >> N_{SN}$; from the relationship given above, the change in temperature due to the narrow line is larger than the change in temperature due to wide line. Since the change in temperature is higher in the smaller line, this line will fail sooner due to electromigration effects, as expected.

Therefore, the standard electromigration median time to failure formula can be rewritten as:

$$MTTF = A(I/A_T)^{-N} e^{[Ea/k(T_0 + \Delta T_N)]} \qquad \textbf{(4.28)}$$

where:

A_T = Cross sectional area of a typical line
I = Current through a typical line

The MTTF can now be determined by establishing the relationship between the temperature and the resistance, along with the previously shown temperature and instantaneous power relationship. Once these relationships have been determined, the tests can be done quickly with a parametric tester that has computer controlled power supplies and digital multi-meters capable of controlling and measuring the instantaneous current and voltage.

Therefore, this test methodology will allow the measurement and application of easily measurable and controllable variables -current, voltage- to apply the necessary stress on interconnection lines and determine MTTF. This requires equipment already available in most semiconductor manufacturing factories. The other advantages of this test are the quick measurements available. Within a couple of minutes, information as to the lifetime of the metal interconnection is readily available.

In practice, the SWEAT tests are designed to show failures in about 30 seconds. Test programs have been developed on parametric testers to do this test rapidly. During the first three seconds, any change in resistance is assumed to be due to Joule heating effects. The power is ramped up fast enough that this is a reasonable assumption. The resistance of the SWEAT structure is checked during this time to insure that there is no overshoot of power. The ramp is for approximately one second and then the current is stabilized for the rest of the three seconds. During the next 27 seconds (approximately), the current is fixed and any change in resistance is assumed to be due to electromigration. Failure is detected when the resistance measurement indicates an open in the SWEAT structure.

The disadvantages of this methodology deal with the fact that there is a very high current density being applied. There have been instances, in the industry, where the parametric testers being used are not fast enough to monitor the breakdown. The Joule heating effect causes the electromigration to run away and it is possible to adjust the instantaneous power to maintain control. This is why the three second ramp rate is used, to prevent overshoot.

Also, the assumption that the test structure under SWEAT stressing behaves the same as the heated wafers used to determine the power/temperature and the resistance/temperature relationship. There can be dynamic heating effects and temperature coupling effects between the wafer and the test structure. This will affect the relationships that have been shown to be linear on the test wafers.

4.4.2 BEM Testing

The second common Wafer Level Reliability test for electromigration that is called the BEM test. BEM stands for **B**reakdown **E**nergy of **M**etal. The similarities in this type of testing to SWEAT testing is that the goal is to use easily determinable external parameters such as current and voltage to measure the lifetime of parts.

Again, the development of the technique starts with the "accepted" electromigration MTTF equation:

$$\mathrm{MTTF} = A^{*}J^{-n}e^{(Q/kT)} \qquad \textbf{(4.29)}$$

Substituting in for the cross sectional area with K*A as shown before with A being the cross sectional area and substituting w (line width) and t (line thickness), and rearranging, Equation 4.29 becomes [14]:

$$\frac{wt}{\mathrm{MTTF}} = K^{*}J^{N}e^{(\frac{-Q}{kT})} \qquad \textbf{(4.30)}$$

The developers of the BEM technique, Dwight L. Cook and Charles C. Hong define a term E to be a materials property which has units of energy per unit length. Firstly they rewrote K* to be:

$$K^{*} = K'\rho(l/v)\sigma \qquad \textbf{(4.31)}$$

Then by substituting in for J and assuming the power of J to be 2:

$$J = I/A = I/wt \qquad \textbf{(4.32)}$$

and combining in the relationship between resistivity and resistance:

$$R = \rho\frac{L}{A} \qquad \textbf{(4.33)}$$

and with some manipulation of the equation, they defined E to be [14]:

$$E = \frac{(wt)^2}{K*} = I^2 \frac{R}{L} MTTF\ e^{(\frac{-Q}{kT})} \tag{4.34}$$

This material property, E, now allows electromigration measurement to take place with variables that can be measured macroscopically such as L, I, R. Since I , T and R are variables during a BEM test and are not constant, a new term such called Median-Energy-to-Fail per unit length can be described as:

$$MEF = \int_{t_0}^{t_{fail}} I^2 \frac{R}{L} e^{(\frac{-Q}{kT})} dt \tag{4.35}$$

They by solving equation 4.34 for MTTF and combining with equation 4.35 yields:

$$MTTF = \frac{MEF}{(R/L)} I^{-2} e^{(Q/kT)} \tag{4.36}$$

Again, the assumption is that the power of J in the standard electromigration equation is shown to be- 2. Cook and Hong then illustrated that the MEF is equal to the MTTF when the current I=1A, temperature T=infinity, and the resistance R/L=1. Essentially, MEF is the MTTF normalized to the previously stated conditions. Therefore, by definition, MEF is independent of the measurement conditions such as current or temperature because it has been normalized.

This means that comparisons between different MEFs are possible without knowing the factors that can not be measured macroscopically. It is easy to measure the MEF of a number of test structures in a lot sample. Each test structure is subjected to a step current. The voltage and the time the current is applied for is recorded. The current is stepped until the metal line fails open. From this the energy-to-fail is easily calculated. Over a number of samples, the MEF is the 50% energy-to-fail distribution point.

To obtain the results in a timely manner, the BEM test must take place in a period of minutes. This means that the current density applied is very high - in the range of 10^7 A/cm^2.

This compares with the hot chuck test method mentioned earlier of 10^5 to 10^6 A/cm. The current is stepped at a rate of typical 5 mA/step with the step duration of 5 seconds. With these settings, the failure of the test structure typically occurs within 20-80 seconds.

Below is a histogram reproduced from the BEM paper by Cook and Hong. This distribution shows the mean current to failure. The current was stepped as described above. This means that a current of 300 mA corresponds to a test time of 90 seconds for the structures used.

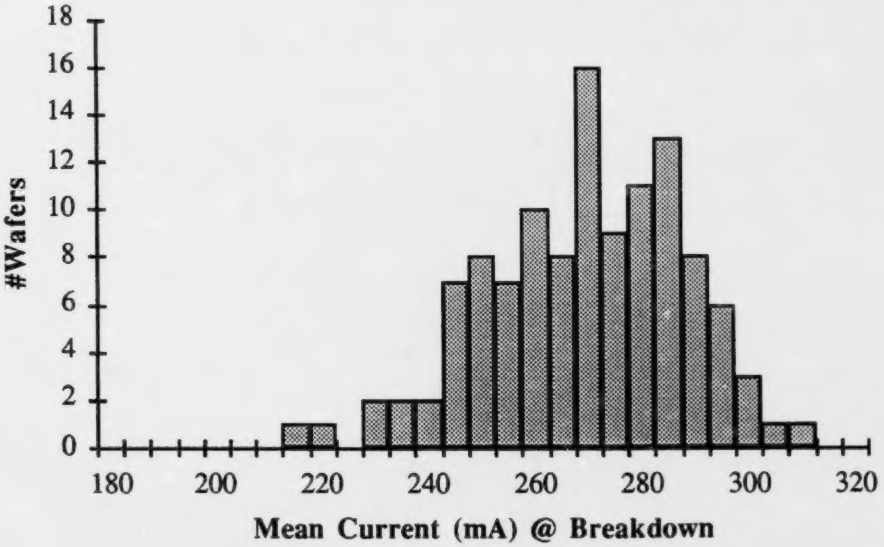


Figure 4.7 Histogram of BEM breakdown currents of production wafers [14].

The BEM technique is similar to the SWEAT technique in that the test structure is stressed by the application of current stress and temperature stress. In both test techniques, the temperature of the test line rises due to Joule heating effects caused by the high current density.

This methodology has the same draw backs as the SWEAT test structure.There is a very high current density being applied. The Joule heating effect can cause the electromigration to run away and can be impossible to adjust the instantaneous power to maintain control.

Also, the assumption that the test structure under BEM stressing behaves the same as the circuits within the actual circuits. The geometry of the test structure is not the same as the circuits. There can be dynamic heating effects and temperature coupling effects between the wafer and the test structure. There will be differences between the test structures and the circuits.

4.4.3 Other Test Techniques

Additional work has been reported in publications on further developments in electromigration test structures. Most of the work has been an extension of the traditional hot chuck method or further work on either the SWEAT or BEM methodologies. Novel test structures have been proposed. Using the test techniques of SWEAT and BEM, other test structures have been designed and tested. Metal interconnect structures such as via chains and contact chains have been used. The effect of step coverage has been explored by doing electromigration testing on a metal line that only goes over underlying structures.

The commonality between all of the various test structures is that the structures attempt to test for structures that occur within actual production circuits. Each of the structures and methodologies are used to try and get a better understanding as to how the lifetime of a circuit is affected by electromigration.

4.5 Electromigration in VLSI Manufacturing

The previous sections have discussed the basic principles of electromigration, what electromigration is and how tests are done to gather circuit lifetime information. Electromigration has been seen to be one of the major lifetime failure modes causing failures in semiconductor VLSI manufacturing today. In manufacturing, electromigration issues have to be dealt with in three separate areas. These are, firstly, in the design of new product and processes, secondly, in the monitoring of the quality of existing product, and thirdly, in the improvement of existing product and processes.

In the first instance, the importance of taking electromigration effects into new product and process design can not be understated. With the advances in technology, the line width of interconnecting conductors has decreased. With this decrease in line width, the current density that the individual conductors has to carry has correspondingly increased. As has been demonstrated earlier, so has the effect of electromigration on circuit lifetime.

With the smaller line width dimensions, electromigration no longer occurs primarily at the grain boundary locations. The surface area of the conductors is now significant when compared to the cross sectional area of the conductor line. This means that electromigration is not only occurring within the grain boundaries but at the surface of the conductors as well. To address these issues, there have been some changes in the structure of the conductors and the materials that make up the conducting lines. To prevent the effect of the surface electromigration, the addition of barrier metals has been brought into manufacturing of VLSI circuits.

In most semiconductor manufacturing processes today, aluminium is the choice for the conductor material used in interconnection. A barrier metal that is used with aluminium is titanium nitride. The conductor line consists of a sandwich structure - aluminium between two layers of titanium nitride. One of the uses of the titanium nitride layer is that the surface of the aluminium layer is bound up and the effects of electromigration along the surface are reduced. This improves circuit lifetime.

Another method that is being used to reduce the effects of electromigration is the addition of alloy material to the conductors. The most common alloy for conductor lines is the addition of copper to aluminium conductors. Empirical experiments have shown that the addition of copper to aluminium of up to 4% by weight have been effective in reducing the effect of electromigration. These experiments have shown clearly that the total mass transport at grain boundaries in aluminium is reduced with the addition of copper [11]. It appears that the addition of copper causes a reduction in grain boundary diffusivity. The exact mechanism is not clearly understood at this time. Further research into the addition of various metals to aluminium is continuing at this time.

Manufacturers are taking advantage of the use of different alloys and barrier metals to increase the product lifetime by reducing the electromigration lifetime failures. It is imperative that testing be done at the design stage of new processes and technologies to minimize the effects of electromigration before the processes are implemented in production. Once the processes have been developed and the effect of electromigration minimized, the next step is to develop an overall quality assurance program with reliability monitoring and evaluation. A large part of the reliability program within manufacturing should be the assurance that finished parts are going to last in the field. This involves rigorous testing of the parts for different failure modes, including electromigration. A program should be implemented to insure that the product does not have electromigration early wear-out problems.

Because SWEAT and BEM use high current densities, in the 10^7 A/cm range, the "standard" MTTF equation does not apply. When the equation was developed, one of the major assumptions was the current density would be in the 10^5 to the 10^6 range. With SWEAT and BEM obviously this is not so. Joule heating effects do take place. Therefore, the MTTF predictions of these test methodologies would not be as accurate as a burn in result.

4.6 Summary

The effect of electromigration is even greater today with the reduced line widths and higher current densities. As has been shown, the Median Time to Failure of a line decreases with the increasing current density within the line. This has meant that the control and testing for electromigration effects in production circuits is more important than ever before. Electromigration testing was accelerated by putting wafers on heated chucks and applying current stress. But this was not fast or practical enough for in line process monitoring so even higher stress accelerated test methodologies such as SWEAT and BEM have been developed.

But these electromigration tests have more inaccuracy due to the higher levels of current density being applied. Joule heating effects become significant. Therefore, the solution is to combine the accelerated testing with the burn in tests and develop relationships and correlation between the two. The manufacturing process can then be base lined and the principles of SPC can be applied to insure manufactured quality in the semiconductor product.

The future of electromigration involves a better understanding of different alloys for interconnection within circuits. It involves further research into barrier metals and multiple metal interfaces within VLSI and ULSI circuits. It will involve a better understanding of the physics of these materials and the electromigration within these materials. Work is ongoing into new test structures and test programs and methodologies. Electromigration monitoring programs are becoming key to manufacturing success in semiconductor technology. A basic overview of what electromigration and how it fits into todays semiconductor VLSI manufacturing technology has been presented.

CHAPTER V
TEST CHIP DESIGN

5.1 Introduction

The semiconductor industry is working on incorporating more Wafer Level Reliability testing into the manufacturing processes to insure higher levels of quality through testing and yet continue to reduce the cost of the final product. This is an almost impossible task. Wafer Level Reliability testing refers to the design of test and test structures that will test for known reliability failures on a parametric tester at the wafer level. To test on a parametric tester, the structures and the tests have to be designed so that the test can be quickly done, usually in under a minute. The Wafer Level Reliability test structures are designed to maximize the acceleration of a single reliability failure mode while minimizing the effect of other yield and reliability failure mechanisms. Wafer Level Reliability test structures are designed for a single reliability failure mode. With Wafer Level Reliability testing, rapid feedback is now possible, instead of the months required for traditional operating lifetime burn in tests. Wafer Level Reliability will not replace the traditional burn in testing for infant mortality and long term life time testing, rather it will supplement this testing. It will provide additional data over a larger sample size that will ensure outgoing quality levels.

The design of a chip consisting of a number of Wafer Level Reliability test structures is discussed in detail in this chapter. Simple structures are used to test for a number of common reliability failures. The individual structures are repeated on a single chip to enable the same test to be repeated within a single test chip. Since most of the Wafer Level Reliability tests are destructive, this allows confirmation of data.

The test chip is unique in that an ASIC base array is used. The ASIC array illustrated is modelled after a standard LSI Logic Corporation Base Array. Until this time, Wafer Level Reliability structures have been full custom designs. The chip is designed to be used within a production environment which may have a number of different base array runs within a single metallization run. This will allow more sampling of data for the reliability of ASIC gate arrays.

5.2 ASIC Manufacturing Requirements

In ASIC Manufacturing, the manufacturing of the semiconductor Integrated Circuit may take place in two parts. The transistor level structures are formed first in the "front end" processes and the interconnecting metals in the "back end" processes. With gate arrays, the "front end" process is like a standard semiconductor manufacturing factory. There are only a limited number of designs that are manufactured. Each of these designs has a different number of gates, with each gate made up of a few transistors. During the design process, the amount of logic gates required will be determined and the appropriate standard "front end" design or base array chosen.

A basic layout of a CMOS base array is again shown in Figure 5.1. Eight transistors are shown, four N-channel and four P-channel devices. The different base arrays essentially consist of various numbers of the basic logic cell shown below. Each different base gate array will have the same logic cell, but have a different number of them from another base gate array within the same technology.

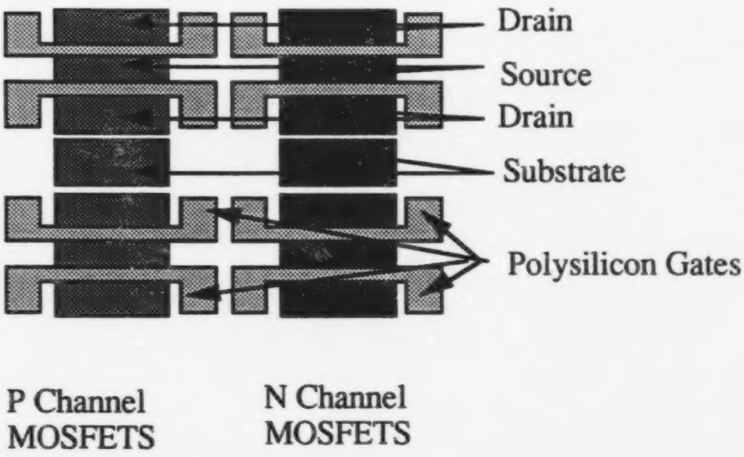


Figure 5.1 Gate array cell

The "back end" processes are where the customization of the gate array circuit occurs. Here, specific metal interconnect layers are defined, according to the specific custom design that is being manufactured. The size of base array was chosen by the amount of logic gates required in the custom design. The function of the logic cell shown above is determined in how the transistors and the transistors of other cells are connected together.

The ASIC Gate array manufacturing can, therefore, take place in two different factories. The base array factory produces the standard base arrays in one and the metallization factory customizes the interconnect. Base arrays are manufactured to go into an inventory that the metallization factory then pulls from. This reduces the manufacturing cycle time. The semiconductors are partially manufactured and in inventory awaiting an order to be finished in the metallization factory. Because of this split manufacturing, a particular order or run of a particular custom design or Application Specific Integrated Circuit -ASIC- may include base arrays made at different times and from different factories. All of this can affect the reliability of the finished product in its final application. This has large implications on the design of a Wafer Level Reliability (WLR) monitor chip.

The Wafer Level Reliability chip must, therefore, be designed such that it will be able to test for the different base arrays. This means that the design will be in the interconnecting metal layers only. The other advantage of this is that the monitoring chip can be placed on the same wafer as product. There are no special processing involved with the manufacture of the monitor chip and therefore, it should see the same processes as the product chip. For the Wafer Level Reliability monitor chip designed in this chapter, it shall consist of four metallization interconnect layers - Transistor contacts, first layer metal interconnect, interlayer dielectric via holes, and second layer metal interconnects. The underlying base array structures will be utilized and connected together to form different Wafer Level Reliability test structures.

Different test structures will be used to test for different reliability failure modes. The three main failure modes tested for are: electromigration failures, dielectric breakdowns and hot carrier failures. These effects have been identified as the major source for most early lifetime wear out failures of semiconductor parts. A set of structures have been adapted to test for these failures. A simple set is used here for there is limited space available on a single chip.

The Wafer Level Reliability structures do not necessarily test for every failure on a semiconductor. The wafer sort and final test routines are designed to test functionality and parametric testing routines are designed to test basic transistor and processing parameters. The Wafer Level Reliability tests are designed to test for specific failures in semiconductors that are not detected by other tests. They will not detect every failure that occurs, but are designed to detect the majority of reliability failures.

5.3 Electromigration Test Structures

Electromigration has become one of the most critical failure phenomenon seen in the early life time failures of today's complex semiconductor circuits. The structures presented in this chapter will deal with different electromigration failures. The standard failures of metal interconnect lines will be tested with the SWEAT structures. The effect of interdielectric processes on metal line lifetimes will be tested through the topology test structure. And the metal connection between layers will be tested through the use of contact and via chain structures.

There will be a number of these structures on the test chip. Electromigration testing tends to be destructive. By having a number of structures within a single chip allows for multiple readings and more data sampling. It allows for testing of the structure, even if a fab defect prevents a particular structure from being used.

5.3.1 SWEAT Structures

To look at the electromigration of an ASIC, the test structures should include all the metal layers. for the ASIC process and layout that is being considered in this design, there are two metal layers. Therefore, there are two separate SWEAT structures, one for the first metal layer and another for the second metal layer.

The second layer metal structure is modelled after the proposed JEDEC standard structure. This is illustrated in Figure 5.2. The structure consists of thin metal 2 lines over metal 1 topology with large heater elements. In the structure, there are twenty to thirty of these elements. With a current stress of 10^{-7} A/cm^2, the structure will only last approximately 30 seconds. The failure will occur in the thin metal 2 lines over the metal 1 heaters. The width of these lines (W) is the same width as the minimum width line in the process. The heater lines are the same width and spacing as the process first layer metal lines. By having the underlying metal lines the same as the process lines, the test structure will be sensitive to the manufacturing process problems.

Empirical results have determined that the best design for the SWEAT structure is having the heat sink large elements 10W long and 10W wide with a narrowing angle of 45°- for processes with dimensions larger than 1 μm in width (W). For sub-micron processes, the recommendation is for 15W for both the length and the width of the large heat sink elements with a narrowing angle of around 70°. This is illustrated in Figure 5.2.

With the metal 2 electromigration SWEAT structure, the proposed JEDEC standard can be adhered to fairly closely. For the metal 1 electromigration structure, there has to be some modifications done due to the fixed base array layout. Because the spacing between transistor groups is not fixed, some of the large heat sink elements will be longer than 10W. This is not the critical part of the design. The test lines at length 10W and width W are the critical elements. They are placed over the existing polysilicon gates such that the polysilicon gates can act as the heater elements. This structure is shown in Figure 5.3.

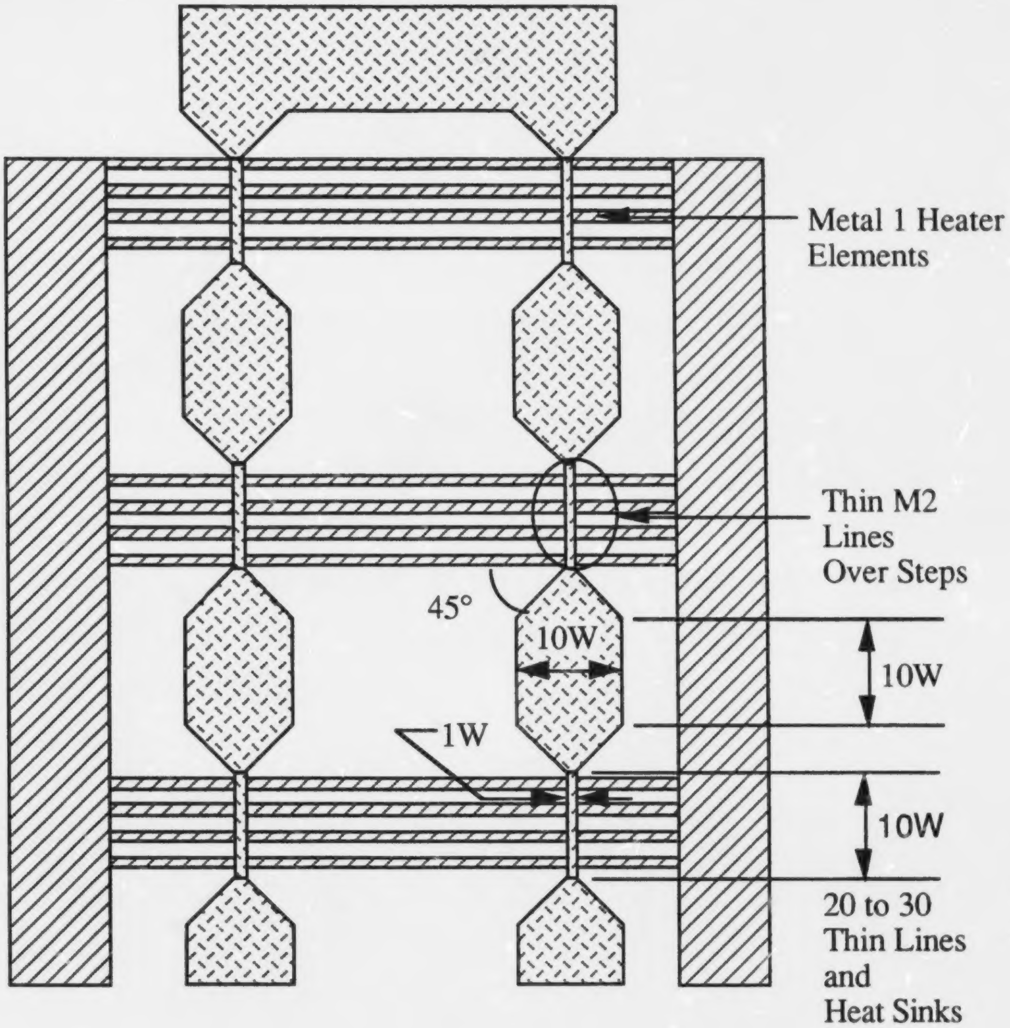

Metal 1 Heater Elements
Thin M2 Lines Over Steps
45°
10W
10W
1W
10W
20 to 30 Thin Lines and Heat Sinks

Figure 5.2 Metal 2 SWEAT test structure

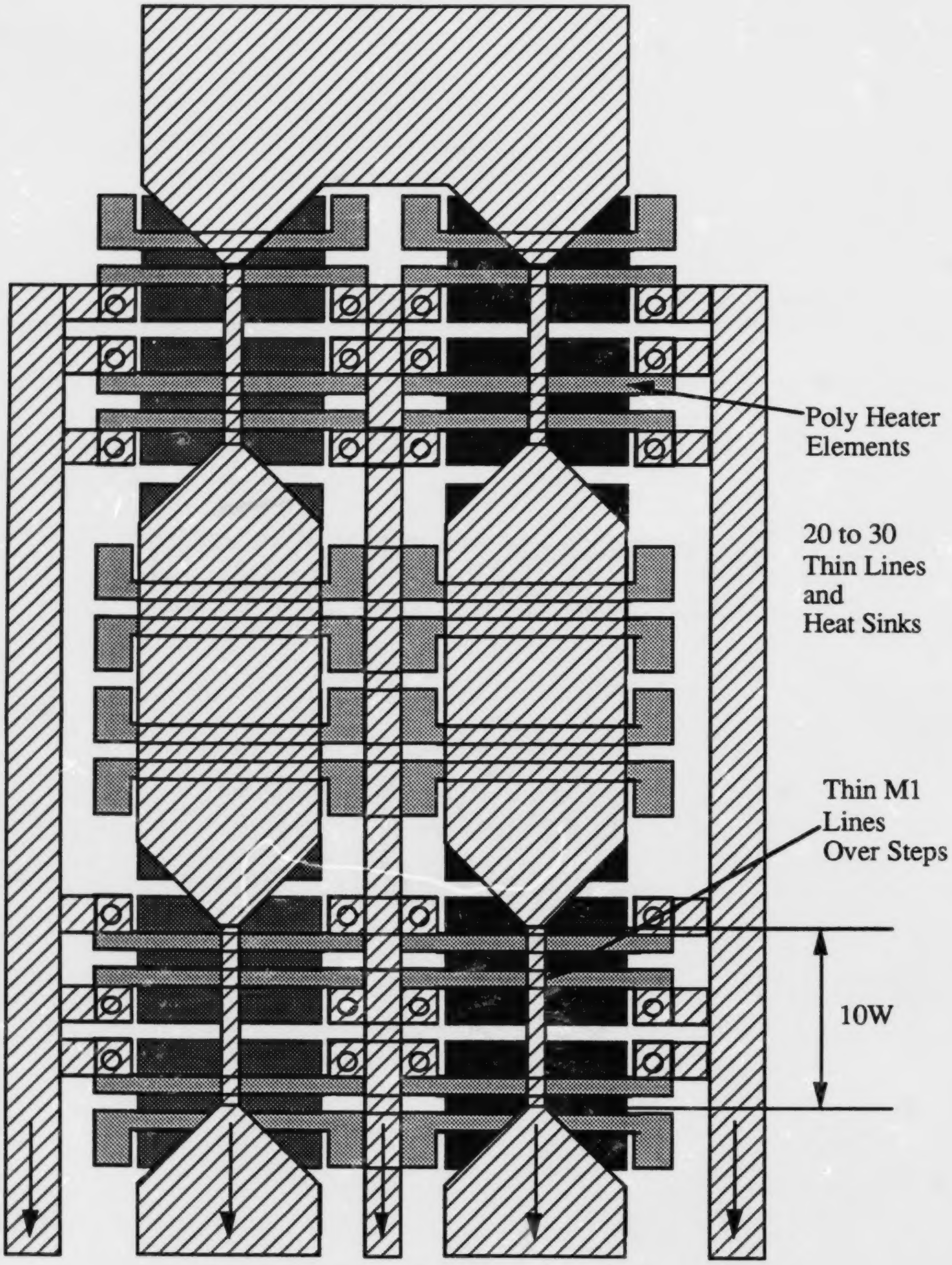

Poly Heater
Elements
20 to 30
Thin Lines
and
Heat Sinks
Thin M1
Lines
Over Steps
10W
To External Pads

Figure 5.3 Metal 1 SWEAT test structure

5.3.2 Topology Test Structure

To determine if there are any problems with the underlying surface topology for the second layer of metal, a topology test structure for electromigration has been proposed by A.S. Oates [15]. In some processes, the step coverage of the metal over the topology is important. Step coverage refers to the percentage of the normal thickness of metal that is in the thinnest portion of a metal line that is over underlying topology or steps. Because of deposition techniques, metal is thinner over side walls of some steps. This thinner metal will have a higher current density than other areas. This higher current density will lead to an earlier failure at the metal line.

To test for the effect of underlying topology, a cross section of what the test structure may look like is shown below:

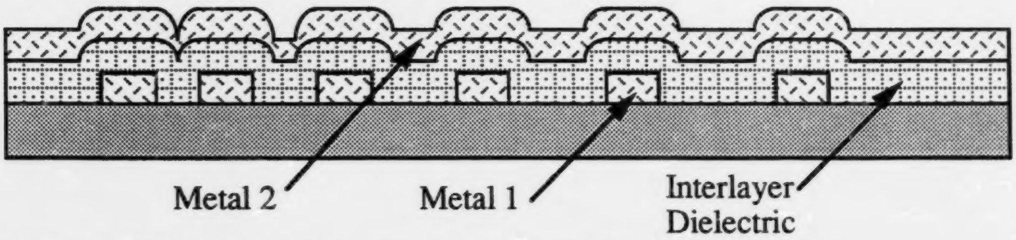


Figure 5.4 Cross section of topology test structure with dielectric topology

The above cross section shows how the test structure may look with a conformal interlayer dielectric. The dielectric shown has the same thickness everywhere when deposited and where there is a underlying feature, such as a metal 1 line, there is a "bump" or "step" in the dielectric. The metal 1 lines of the test structure are not evenly spaced. The spacing is done logarithmically. This is because there will be possibly a spacing between metal lines where the process will give a low step coverage. If the metal 1 lines are very close together, then there is essentially no step and, therefore, no loss in metal 2 line thickness. If the lines are very far apart, there is no large effect on deposition and the step coverage is good. It is in between these spacing where trouble can occur. As shown in the diagram previous, the first two metal 1 lines are close together. The metal 2 line covering the dielectric at that point has a very sharp peak to it, a cusp. This area is prone to stress and cracking. Here failure can occur because the metal is thinned at this point due to the stress pulling the metal apart, or due to the lack of metal in the "cusp".

The spacing between the second and third metal 1 line can also show problems in the metal 2 step coverage. Most deposition techniques for metal today involve sputtering techniques, where the atomic metal follows a line of sight from the target to the wafer. Close together features tends to cause a "shadowing" effect which reduces the amount of metal deposited. Higher steps or features also has this effect. The net result is less metal deposited and, therefore, a lower step coverage. This leads to areas of potential electromigration failures.

The height of steps and the amount of topology depends on the process in question. The figure below illustrates 100% step coverage due to a fully planarized process.

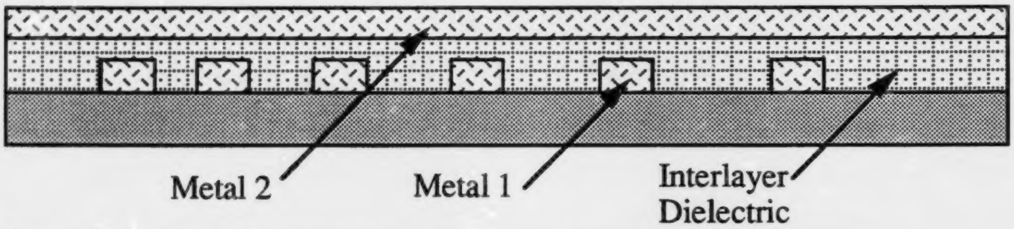


Figure 5.5 Cross section of topology test structure with no dielectric topology

Here there is no thinning of the metal 2 lines due to steps. But a fully planarized process may not be required or even desired. It can have its own yield and reliability problems. Most VLSI and ULSI processes have some degree of planarization in the dielectric processes. The test structure proposed is shown in Figure 5.6.

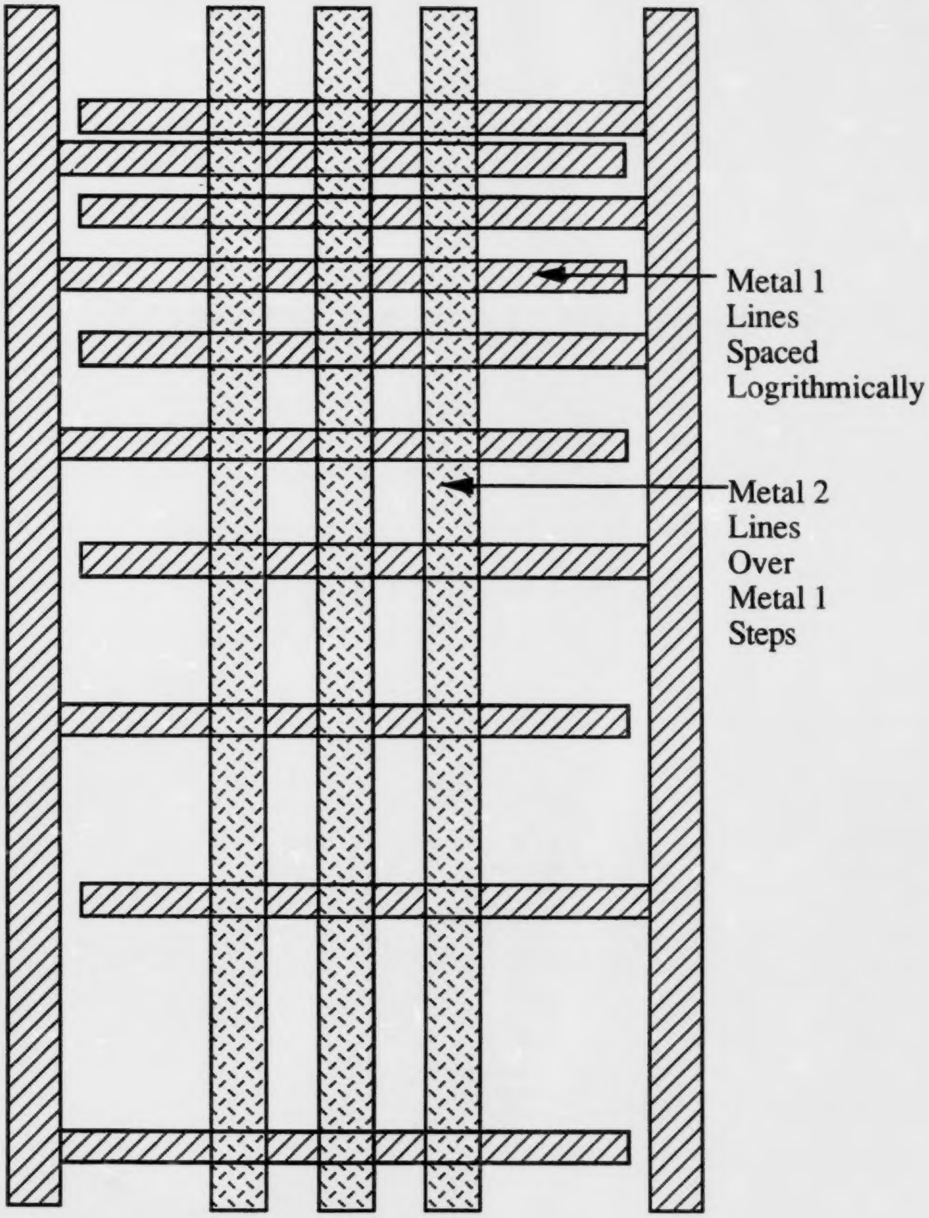

Metal 1
Lines
Spaced
Logrithmically
Metal 2
Lines
Over
Metal 1
Steps

Figure 5.6 Topology electromigration test structure [15]

The structure has three metal 2 lines that go over the metal 1 lines. The metal 1 lines are spaced logarithmically and are connected alternatively to one of two metal 1 buses. With certain spacings of the metal 1 lines, there can be deep narrow groves formed in the covering interlayer dielectric. There may be excess metal 2 in these groves that does not get etched out due to the photolithographic processes or etch processes. These left over metal 2 in the deep narrow groves or cusps are referred to as stringers. These stringers can cause yield problems and this test structure, therefore, serves a dual purpose - both as a wafer lever reliability structure and a yield monitoring structure for the planarization related processes. On the test chip, the test structure is repeated several times - one structure connects to the next structure. A series of the topology structures span the entire length of the test chip.

5.3.3 Contact and Via Test Structures

The last electromigration test structures used are standard contact and via chains. A chain consists of a number of metal interconnections made through a dielectric connected together in a single string or chain of contacts or vias. As with the topology test structure, step coverage is again one of the issues. To make contact with the underlying transistors, contact holes are made in the protective dielectric. (Usually BPSG). Metal 1 then is able to make electrical contact with the underlying transistors. With some processes, the metal 1 is deposited directly into the contact holes. Here, because of the small dimensions, there is again a thinning in the metal that is in the contact itself. This is the location where an electromigration failure can take place.

An example of a cross section of a contact chain is shown in Figure 5.7 The metal 1 is thinner in the contact holes because of the difficulty of depositing onto side walls of the contact holes. The contacts are a N+ chain. For P-channel transistors, there is a P+ chain. Contacts to polysilicon make up the third contact chain.

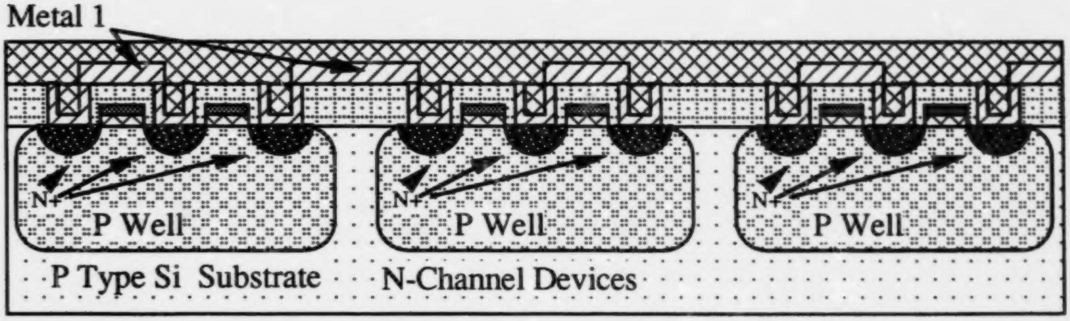

Metal 1
N+
P Well
N+
P Well
N+
P Well
P Type Si Substrate
N-Channel Devices

Figure 5.7 N+ contact chain cross section

An additional electromigration failure can occur in the silicon contacts. If there are process problems with the deposition of the metal 1 or too high temperature during processing, a phenomenon known as contact spiking can occur. Most metallization in VLSI is an alloy of aluminium. Usually the aluminium alloy has silicon in it because aluminium will naturally adsorb silicon. If there is too much heat or too little silicon in the alloy, the silicon will migrate out of the N+ or P+ contact into the aluminium. Over time, electrical current will enhance this effect. The metal will then migrate through the junction (N+ or P+) consuming the silicon. Once through the junction, the transistor is shorted out, and the device stops working. The spiking effect is illustrated in Figure 5.8.

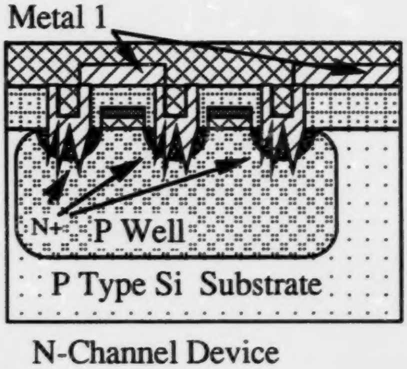

Metal 1
N+
P Well
P Type Si Substrate
N-Channel Device

Figure 5.8 Metal spiking in N+ junctions

N+, P+ Contact Chain and Polysilicon Chain test structures are shown in Figures 5.8 and 5.9, respectively.

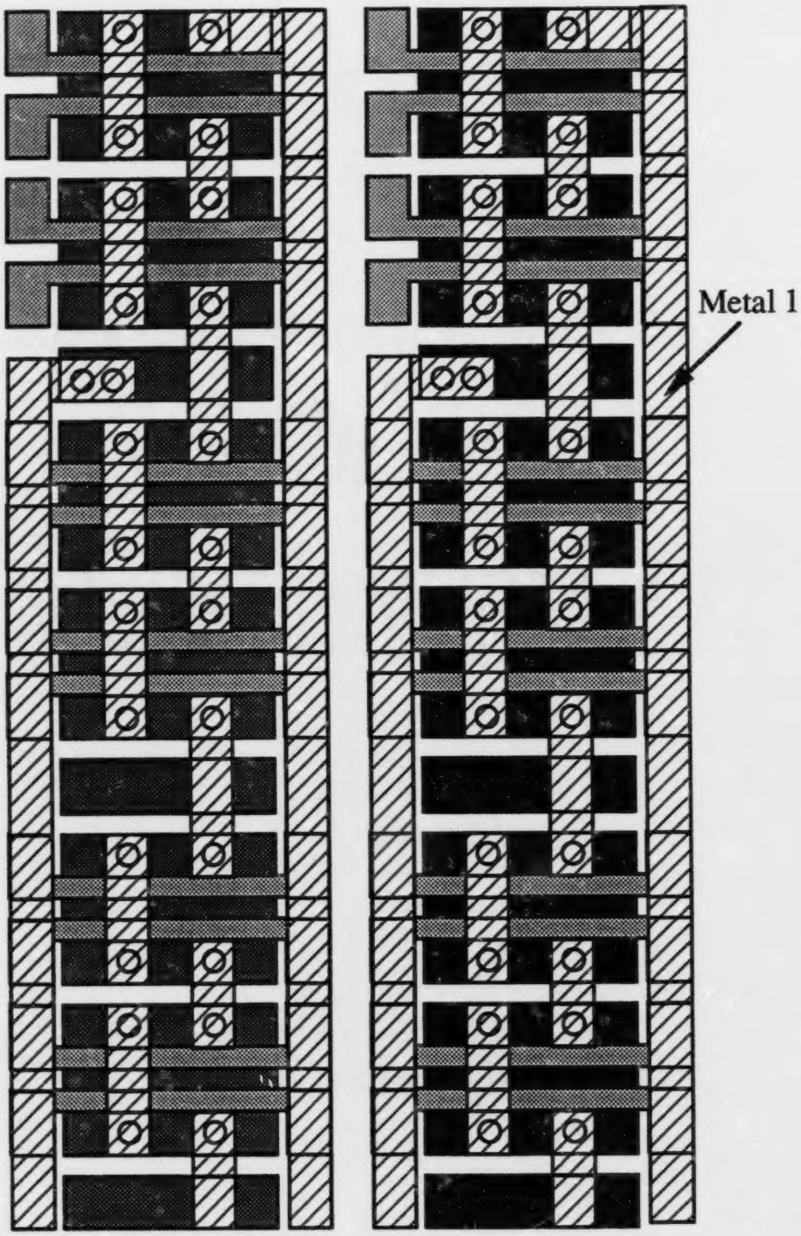


Figure 5.9 N+, P+, contact chains test structures

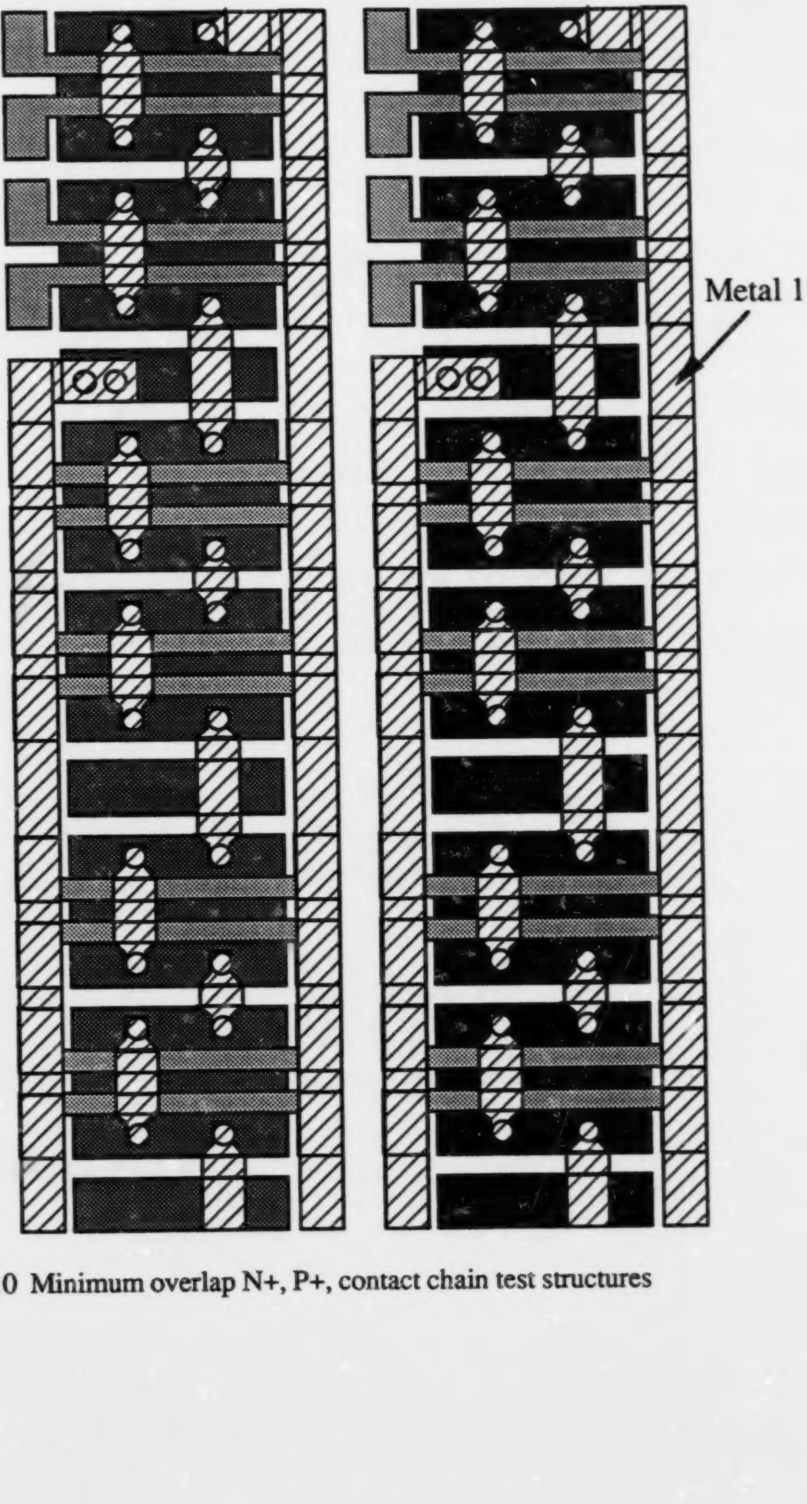


Figure 5.10 Minimum overlap N+, P+, contact chain test structures

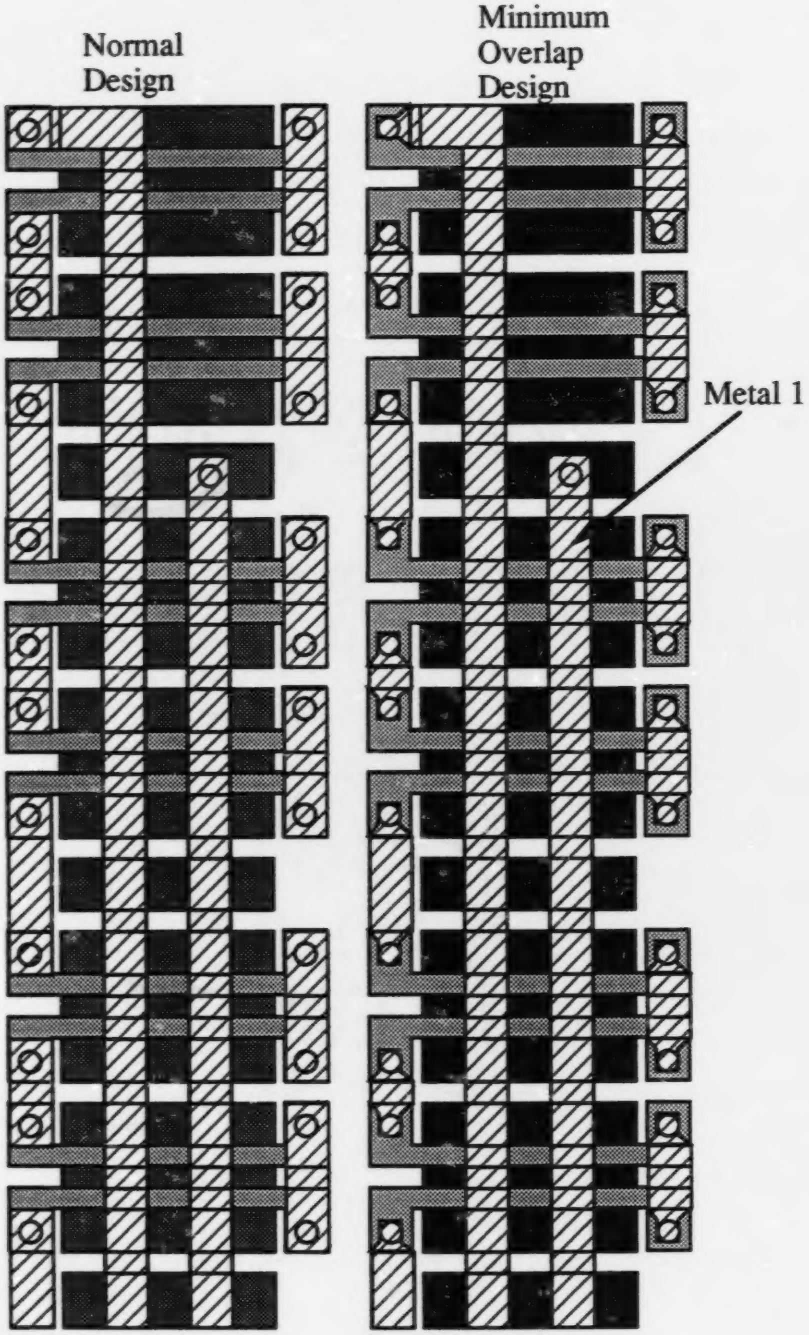

Normal
Design
Minimum
Overlap
Design
Metal 1

Figure 5.11 Polysilicon contact chain test structures

The via chain test structure is similar to the contact chain structures in that the structure is a chain of via contacts between the metal 1 layer and the metal 2 layer, with the vias being the holes formed in the interlayer dielectric. This is illustrated below:

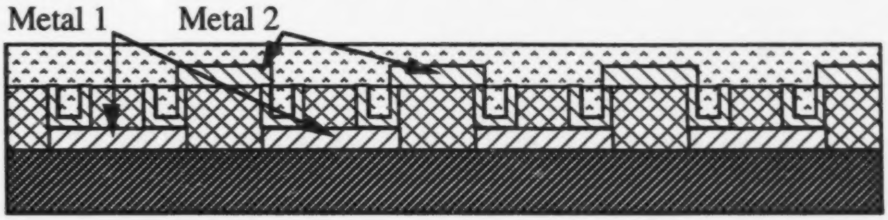


Figure 5.12 Via chain cross section

Here, the main reliability concern with some processes is, again, with step coverage issues. The metal may thin because of deposition in the vias, causing areas were electromigration failure will occur first. All the chain structures in the test chip will consist of only a hundred or so vias or contacts. Because most accelerated electromigration testing uses electrical stressing for heating, the overall resistance of the chain must be kept low to obtain current densities such that the Joule heating effects of the metal do not cause the failure instead of the electromigration effects.

A via test structure is shown in Figure 5.13.

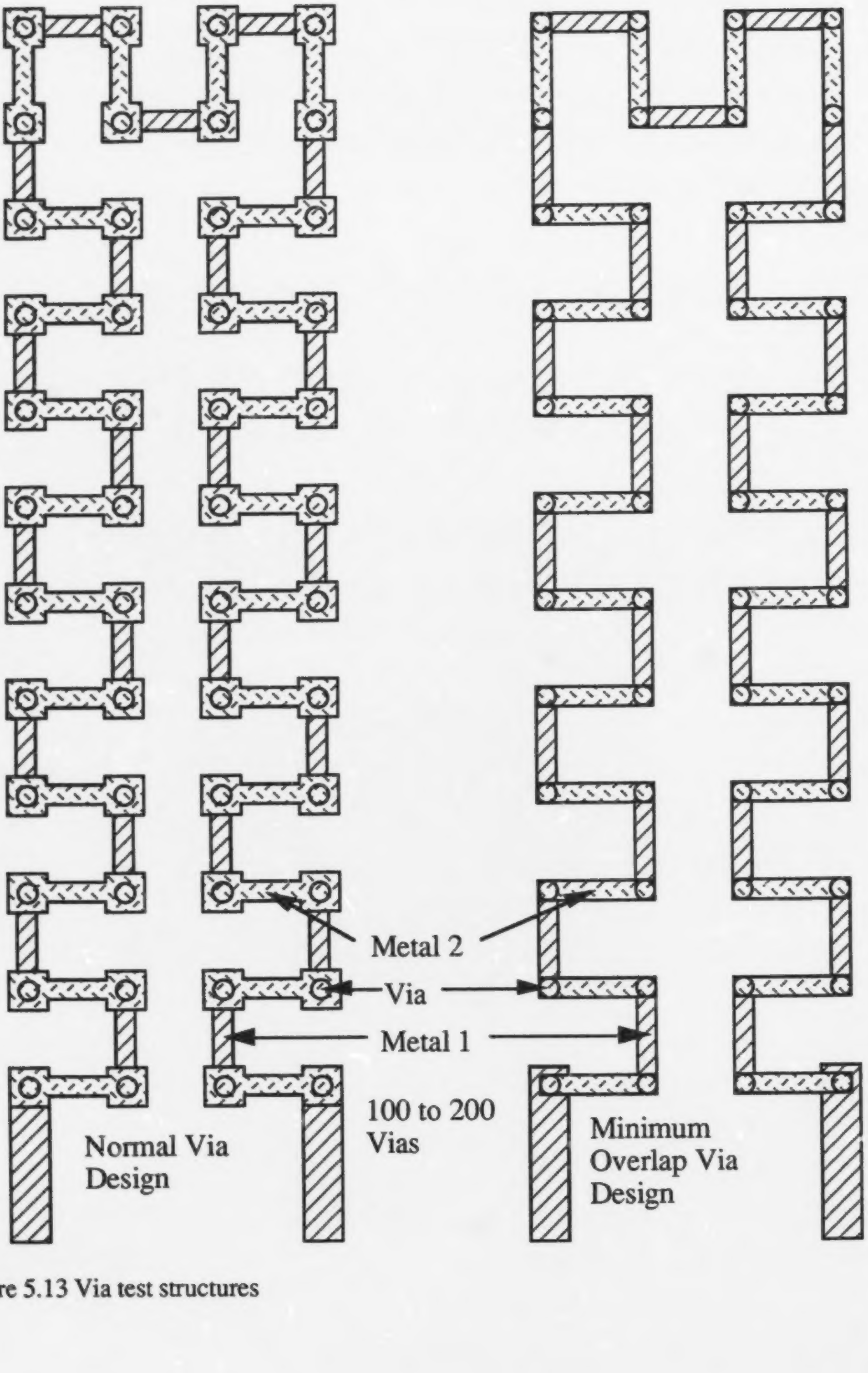

Metal 2
Via
Metal 1
100 to 200 Vias
Normal Via Design
Minimum Overlap Via Design

Figure 5.13 Via test structures

5.4 Dielectric Breakdown Test Structures

The other subsets of structures deal with the early failures of circuits due to dielectric failures. During manufacturing, high energy electrons and ions can get trapped in the dielectric, weakening the dielectric. In MOS circuits, this is critical. These trapped charges can provide a path for electrons and weaken the crucial gate oxide. Repeated on/off stressing will cause the oxide to breakdown at a lower than operating voltage, causing a reliability failure.

To test for these reliability failures, wafer level tests have been devised. The tests consist of stressing the dielectric with either current or voltage stresses. The structures used for these tests are large capacitors. The reason for the use of large capacitors is to take advantage of the greater area of dielectric used than in the actual circuit. The greater the area, the more chance there is of point defects. By testing a large area of dielectric for breakdown, the more accurate the data concerning the reliability of the dielectric will be.

As stated before, the dielectric is stressed through either voltage or current electrical stressing, until the dielectric breaks down. Constant current has been used to measure the charge to breakdown, or QBD (Q=Charge, BD=Break Down). A constant current is applied to the dielectric capacitor, and the voltage drop is measured. When the voltage across the capacitor drops to zero, the dielectric has broken down. By knowing the amount of constant current and time to break down, the QBD can be easily calculated. The longer it takes for the dielectric to break down, the better the dielectric is.

For wafer level testing, the time to do a QBD test can be a number of minutes. For production testing, other methods have been used to speed up the QBD test. These methods have included a ramp or a stepping increase of the current. When the current is ramped or stepped, the test is referred to as JRAMP. The drawback with these methods is that if the test is done too fast, there is a loss of resolution and a potential loss of data.

The second method involves setting up a constant electric field. By applying a constant voltage, the current through the capacitor is monitored. When the stress is large enough, the dielectric will break down and the current will increase. Again, the test may take a number of minutes. By ramping or stepping increasing the voltage, the test can be done faster. This ramped or stepped test is referred to as VRAMP. Again, the trade off is resolution. If the test is done too fast, there may not be any difference between good dielectrics and weak dielectrics.

The structures used in the test chip are capacitors. The capacitors are formed by using existing structures of the base logic array. Three structures are used to test the reliability of the different dielectrics. The dielectrics tested are: the interlayer dielectric, the BPSG oxide on top of the transistors, and the gate oxide.The gate oxide is the critical dielectric that effects the performance of the switching logic transistors. The test circuits that are used are shown in Figures 5.14, 5.15 and 5.16. In the diagrams, the meters are multimeters and the supplies can be either current, voltage, AC or DC supplies.

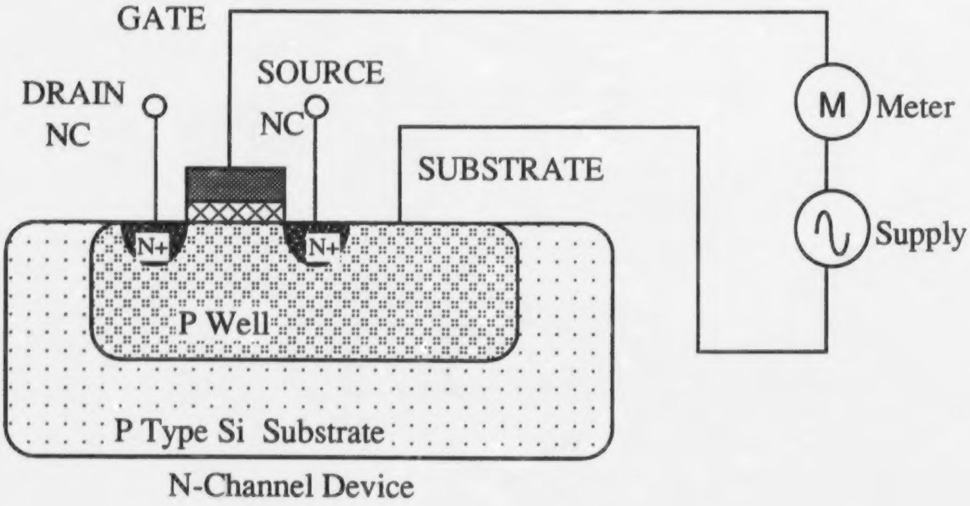

GATE
DRAIN
NC
SOURCE
NC
SUBSTRATE
M
Meter
Supply
N+
N+
P Well
P Type Si Substrate
N-Channel Device

Figure 5.14 Gate oxide test circuit

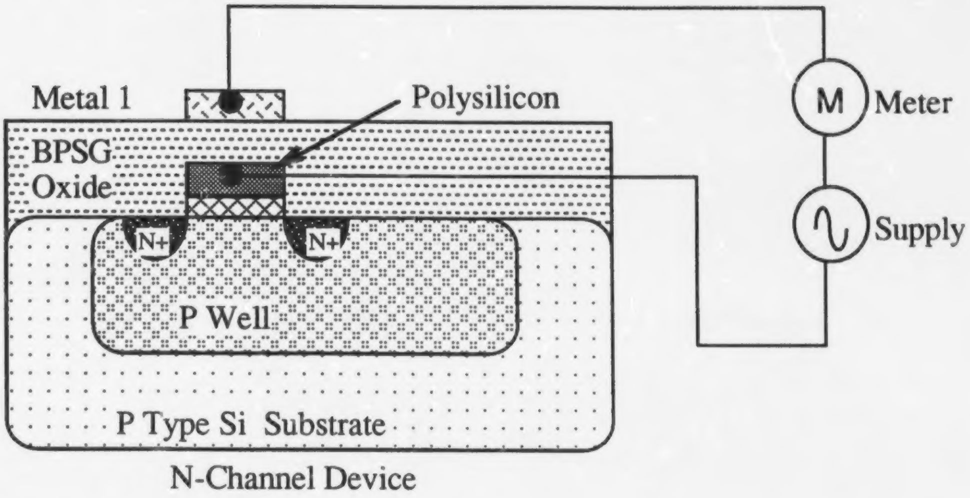

Metal 1
Polysilicon
BPSG
Oxide
N+
N+
P Well
P Type Si Substrate
N-Channel Device
M
Meter
Supply

Figure 5.15 BPSG oxide test circuit

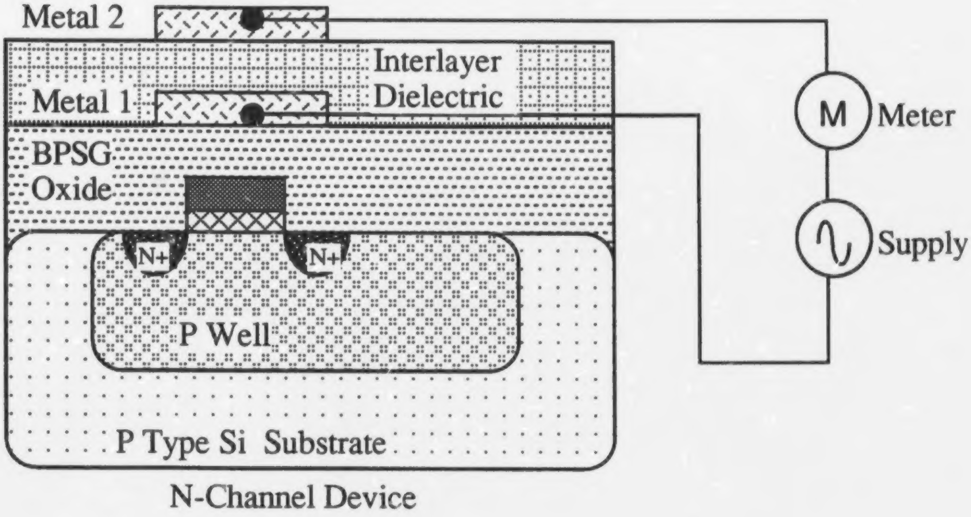

Metal 2
Interlayer
Dielectric
Metal 1
BPSG
Oxide
N+
N+
P Well
P Type Si Substrate
N-Channel Device
M
Meter
Supply

Figure 5.16 Interdielectric test circuit

The test circuit has three structures that represent the above circuits. The circuits are in a cellular approach, similar to the ASIC circuits. For the gate oxide test circuits, the large capacitor is formed by the connection of many small capacitors. The substrates are connected together as well as many polysilicon gates. The same type of design applies with the BPSG oxide. Many polysilicon gates are connected together and metal 1 structures over top of the gates are connected together.

The interlayer dielectric capacitor is formed by making a large capacitor directly. The reason that this is possible is that the ASIC gate array involves the customization of the metal interconnect layers only. The other dielectric structures described earlier involved layers that made the basic transistors. These "front end" layers are defined earlier in the process. With an ASIC Wafer Level Reliability test chip, these layers are previously manufactured. The test circuit is formed by using the metallization layers to connect together the "front end" layers to form the test structures. Figures 5.17 and 5.18 show the design of the dielectric test structures.

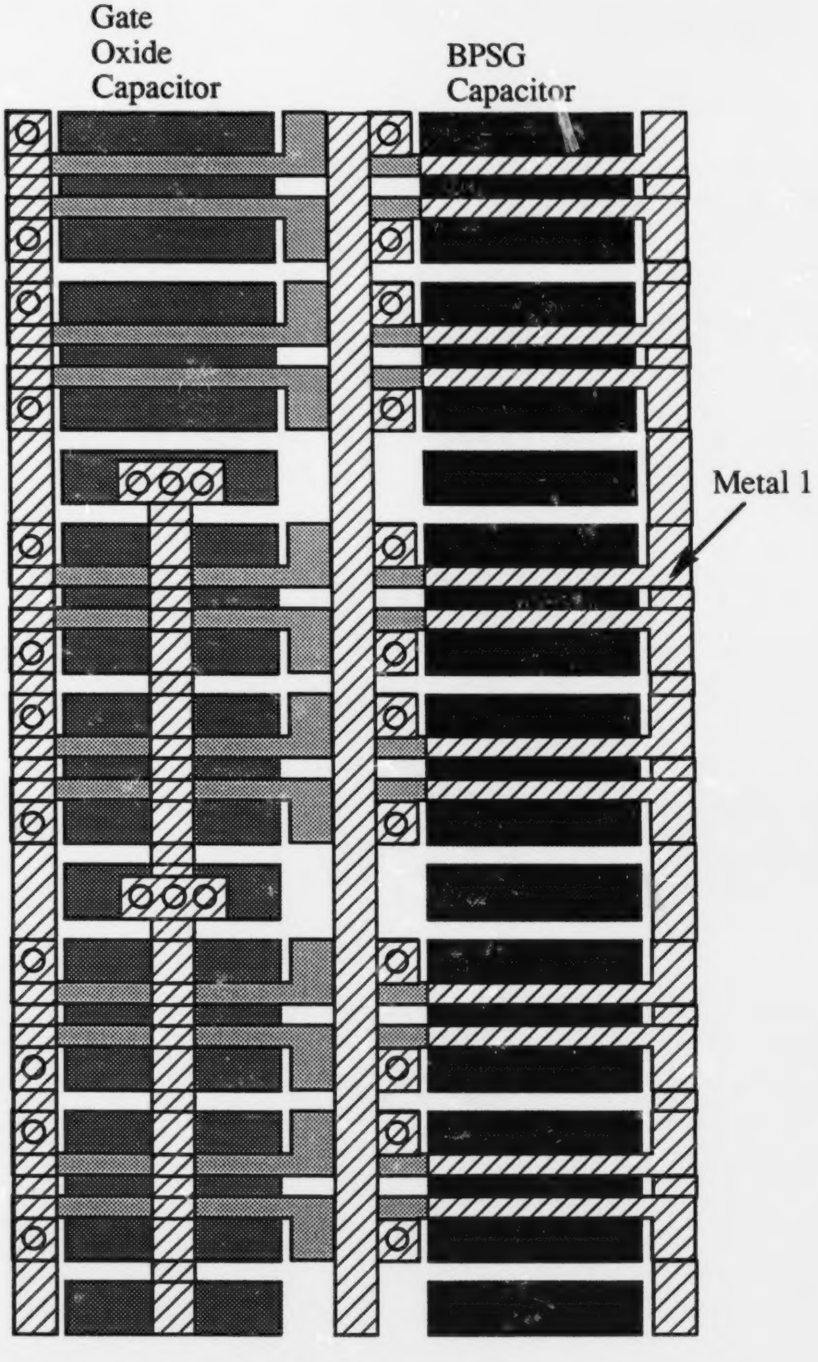

Gate
Oxide
Capacitor
BPSG
Capacitor
Metal 1

Figure 5.17 Gate oxide and BPSG test capacitor structures

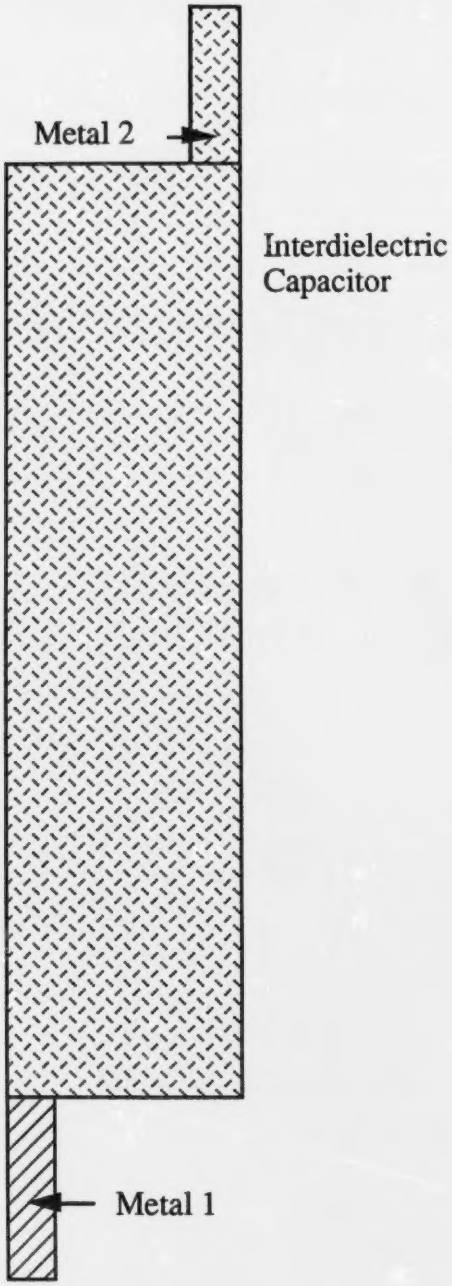

Metal 2
Interdielectric
Capacitor
Metal 1

Figure 5.18 Interdielectric test capacitor structure

5.5 Hot Carrier Test Structures

The third type of common reliability failure is the failure of transistors due to the injection of hot carriers into the insulating oxide. Hot carriers are trapped electrons or holes within the sensitive insulators within the transistor structure. In CMOS, they are either holes or electrons trapped in the gate oxide. A good way to illustrate this phenomenon is to show how the hot carriers get trapped in the gate oxide of an N channel MOS transistor. The transistor in the example shown below has the substrate and source grounded and the gate and drain at +5 Volts. The transistor will be in the on state and electrons (hot carriers) will be travelling from the source to the drain as shown in Figure 5.19 below.

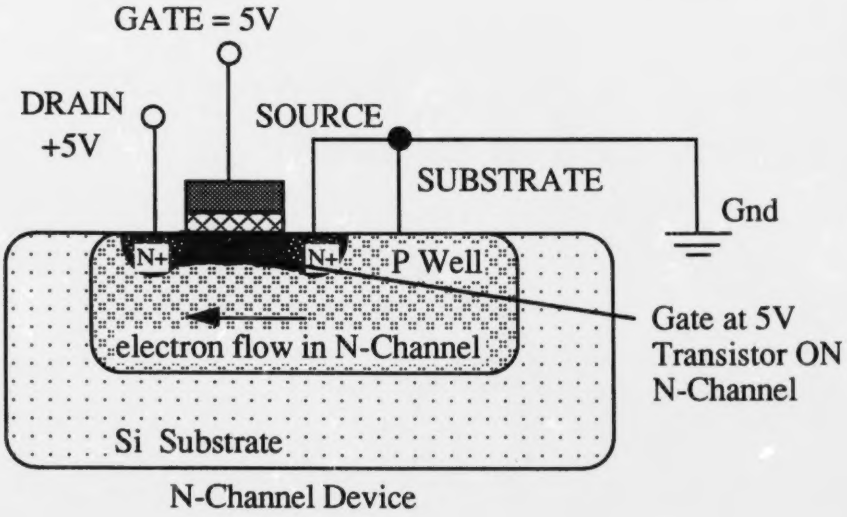


Figure 5.19 Electron flow in N-Channel MOS device

The electrons travel in the N-Channel that is formed by the biasing of the transistor shown above. This is the classic model of how a simple MOS transistor functions. However, the energy of individual electrons in a lattice structure is not constant. Some electrons travelling through the N-Channel in the above example will have more kinetic energy than other electrons. At a higher potential voltage applied as a bias to the transistor, there is statistically more electrons at a higher energy level.

Basic semiconductor physics have shown that the energy for a carrier to surmount the Si-

SiO_2 barrier is 3.1 Volts [16]. Because most CMOS ICs have a bias voltage of +5 Volts on the gate, there are some carriers that will have enough energy to surmount the Si-SiO_2 barrier. There will be some electrons in the N-Channel that will travel through to the gate contact, but some electrons will be trapped in the gate oxide lattice structure. The term "Hot" refers to those electrons that have sufficient enough energy to surmount the barrier. The probability of this occurring increases with the potential applied as the bias voltages. This is shown in Figure 5.20 below.

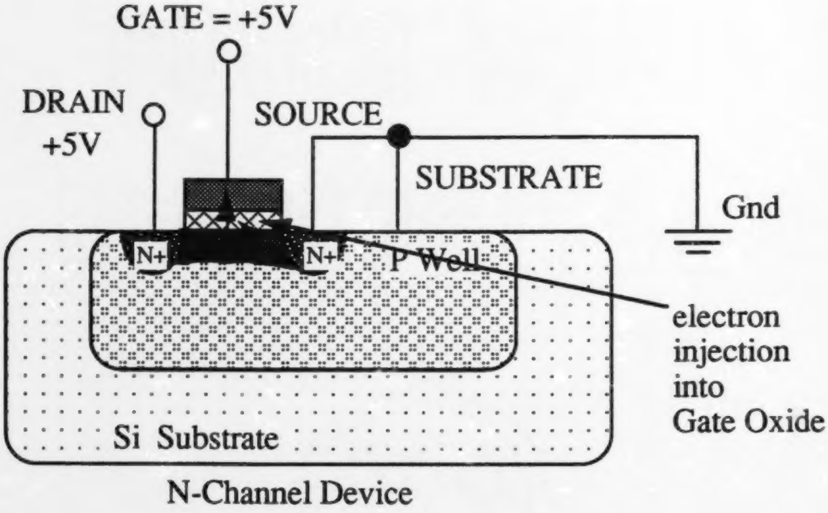


Figure 5.20 Hot electron injection into N-Channel MOS gate oxide

The insulating properties of the gate oxide is affected by the introduction of extra electrons in the lattice structure. Over time the amount of electrons trapped in the gate oxide can cause threshold voltage instabilities and even gate oxide leakage and breakdown. This can lead to part failure.

Hot Carrier failures are becoming more of a concern within the semiconductor industry. With the constantly shrinking geometries, the gate oxides and other sensitive insulators are becoming smaller and more sensitive to trapped charges. The example given was of hot electron injection into a gate oxide of an N-channel MOS transistor. Similar effects occur with positive charge (holes) injection into P-Channel devices. This effect is also not limited to MOS or CMOS transistors alone. Bipolar devices can experience failure due to hot carriers being trapped between insulators that separate transistors or emitter-base regions.

The failure that occurs in MOS transistors is characterized by the degradation of measurable transistor parameters. The change in the transistor parameters is an indication of the hot carrier injection. The lifetime of these semiconductor devices has often been defined by the MOS drain to source current (I_{DS}) decreasing by 10% [17]. There is a degradation of I_{DS} because with more impurities in the gate oxide, the leakage current ($I_{Gate\text{-}Substrate}$), increases. From this, the time dependence of N-channel MOSFET transistors has generally been given as [17]:

$$\log\left(\frac{\Delta I}{I}\right) = K^{*}\log(I_B) + M^{*}\log(T) + K_0 \tag{5.1}$$

where:

$\frac{\Delta I}{I}$ = fractional change in the drain current in the transistor linear region

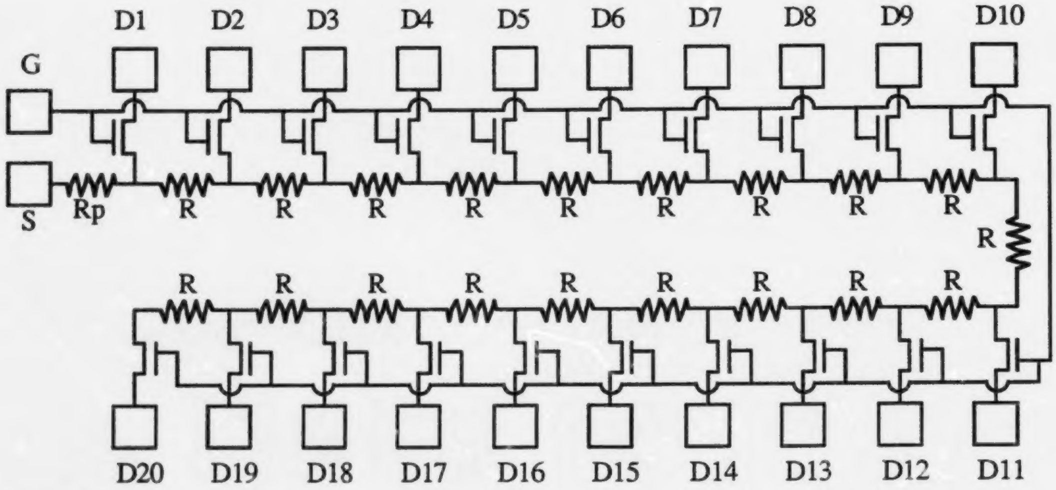
I_B = gate-substrate current during stress

T = time

K,M,K0 = constants for a particular design and channel length.

With the shrinking geometries in semiconductors, the effect of hot carrier degradation on transistors lifetime is increasing. The gate widths are decreasing causing a higher current density and therefore, more carriers with higher energy. The gate oxides are also thinner. With less area in the transistor, the effects of hot carriers become more pronounced.

A novel test structure to show hot carrier degradation of MOS transistors has been proposed by Thomas Kopley [18]. This structure consists of twenty identical transistors with a common source and a common gate. Each of the drains are connected to external pads. The circuit is shown in Figure 5.21 below.



Legend:
D=Drain
S=Source
G=Gate
R=Interconnection Resistance

Figure 5.21 Hot carrier test circuit

The circuit used above is used to observe changes in the transistor parameters. The procedure for looking at the hot carrier injection is outlined below:

1. Use FETs as voltage probes to measure R of the common source
2. Measure IV curve
3. Stress the structure using some FETs for voltage measurements
4. Calculate V drops and lifetimes.

By knowing the R of the common source, then the substrate currents can be calculated before and after. This structure uses some of the transistors as voltage probes, typically 1 out of 5, and these are not stressed. The stress is electrically with higher voltages and currents. From the test the degradation of the substrate currents can be calculated and the corresponding lifetimes.

This structure has a number of advantages. Up to 20 transistors can be used for lifetime vs. V_{DS} tests for only one stress. Errors in individual transistors tend to average out. The FETs can be used as voltage probes to determine the voltage drops IR. In CMOS base arrays, there would be two similar structures, one for N-Channel transistors and one for P-Channel transistors. The structure for both is shown in Figure 5.22. This test structure was first presented by Thomas Kopley at the 1991 Wafer Level Reliability Workshop in October 1991 [18]. The N-channel structure is identical to the P-Channel structure, except that the different type of transistors are used.

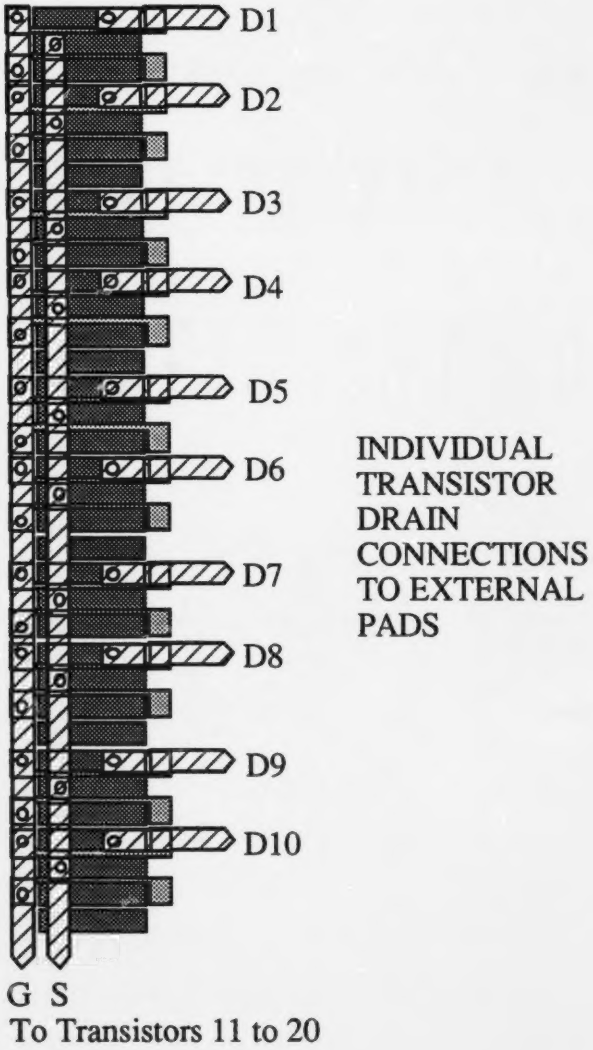

D1
D2
D3
D4
D5
D6
D7
D8
D9
D10
INDIVIDUAL TRANSISTOR DRAIN CONNECTIONS TO EXTERNAL PADS
G S
To Transistors 11 to 20

Figure 5.22 Hot carrier test structure

5.6 Test Chip Layout

The structures that have been shown previously have been designed to be part of a single test chip. Because most of the tests at the wafer level tend to be destructive, the various test structures are repeated. This allows for more data and helps eliminate wafer processing yield defects from reliability defects. The chip is designed to be fabricated along side actual product chips on a base array. This allows for real time collection of reliability data on a wafer. This is some times referred to as "test sites" on a wafer. Instead of a product chip, a test chip such as the Wafer Level Reliability test chip described here would be fabricated. Usually, there would be a few test sites distributed across a wafer. This is illustrated below:

With a few sites on production wafers, it is therefore important to have more than one of a single type of test structure on the Wafer Level Reliability test chip. The other feature of the structures designed is that the structures have their inputs and output pads next to each other. When probing the test chip manually, this makes for ease of testing. The exception to this is the topology test structure. Here, a number of structures are connected end to end across the entire chip. An example of how the chip may be configured is shown in Figure 5.23. An actual ASIC gate array may have more or less pads, dependent upon the size of the array. The example is for illustrative purposes, and is not to scale.

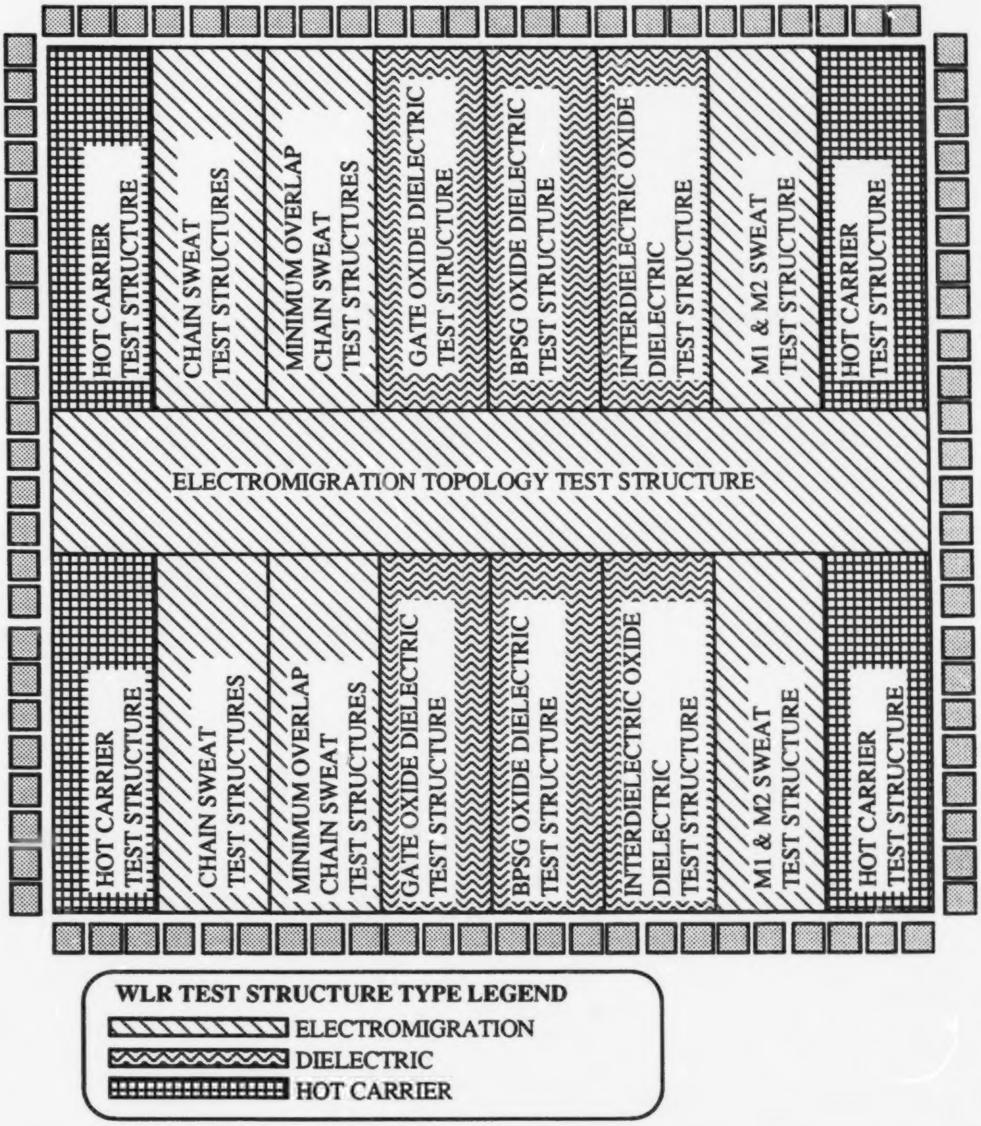

HOT CARRIER TEST STRUCTURE
CHAIN SWEAT TEST STRUCTURES
MINIMUM OVERLAP CHAIN SWEAT TEST STRUCTURES
GATE OXIDE DIELECTRIC TEST STRUCTURE
BPSG OXIDE DIELECTRIC TEST STRUCTURE
INTERDIELECTRIC OXIDE DIELECTRIC TEST STRUCTURE
M1 & M2 SWEAT TEST STRUCTURE
HOT CARRIER TEST STRUCTURE
ELECTROMIGRATION TOPOLOGY TEST STRUCTURE
HOT CARRIER TEST STRUCTURE
CHAIN SWEAT TEST STRUCTURES
MINIMUM OVERLAP CHAIN SWEAT TEST STRUCTURES
GATE OXIDE DIELECTRIC TEST STRUCTURE
BPSG OXIDE DIELECTRIC TEST STRUCTURE
INTERDIELECTRIC OXIDE DIELECTRIC TEST STRUCTURE
M1 & M2 SWEAT TEST STRUCTURE
HOT CARRIER TEST STRUCTURE
WLR TEST STRUCTURE TYPE LEGEND
ELECTROMIGRATION
DIELECTRIC
HOT CARRIER

Figure 5.23 Test chip layout

For the purposes of clarity, the test structures designed in this paper show only part of the entire structure. For example, the Via chain structure that is shown is repeated until the required number of Via contacts are realized. The SWEAT structure consists of 20 to 30 individual thin lines of width W and heater elements. These are the same type of design concepts that are used in ASIC designs today. The structures illustrated in the figures are like individual cells in an ASIC CAD library. The same cell structure is repeated until the desired test structure is realized, much like individual gates and cells are connected together to form higher logic functions in ASIC logic gate arrays.

5.7 Summary

Basic structures to test the three main Wafer Reliability failures have been designed to be fabricated on an ASIC base array. In the past, these structures were only designed for full custom applications. This unique design allows for real time reliability testing in an ASIC production facility. The structures are designed to make use of the underlying transistors in the realization of the test structures. They are designed to fit on top of an existing gate array. This allows for the use of the structures as test sites in a production ASIC environment.

The structures are designed to test for three of the most common reliability failures in ASIC semiconductors today - electromigration failures, dielectric failures and hot carrier failures. There are a number of different test structures within these areas to test for different process failures, i.e topology electromigration structure has been designed to test for planarization reliability process problems.

The number of types of structures has been kept to a minimum of about ten different structures. Wafer Level reliability is a fairly new field. Research is ongoing into new structures and variations on the structures presented here. The goal of this design was to allow for data collection and research into reliability of an existing product line, not into new test structures.

CHAPTER VI

BURN IN RELIABILITY

6.1 Introduction

Reliability is a major concern in the manufacture of complex semiconductor components. Manufacturers usually have reliability monitoring programs to insure that semiconductor Integrated Circuits (ICs) are reliable. A major part of these programs is burn in. Burn in refers to the accelerated testing that is done on a sample of components to determine the product life time in the field. Burn in is the traditional method that semiconductor manufacturers use to monitor the reliability of the product they manufacture. Burn in refers to a series of tests designed to stress the parts and precipitate the failures. By stressing the parts, the lifetime of the parts is reduced to a time where it is practical to gather data about the reliability of the product. This stressing accelerates the effects of the failure mechanisms. Through use of models and relationships such as Arrhenius equation, information about the actual lifetime of the semiconductor component in the field is estimated.

Within the field of semiconductor manufacturing, Burn in has come to include all the accelerated stressing used to monitor reliability. Burn in comes from the fact that most of the stressing is by the use of temperature stress, hence the term, "burn in." Not all the stresses that are used to shorten the lifetime of tests lots involve temperature. However, the High Temperature Dynamic Lifetest is the one test that is used throughout the industry to predict the life of components.

Burn in is used to identify problems with existing manufacturing processes, technologies, designs and product. Most semiconductor manufacturers use a constant sampling methodology. A number of ICs are subjected to accelerated testing to determine lifetime, failure rates, and failure mechanisms. These are actual parts that were sampled from ongoing manufacturing. Any parts or components that fail during the accelerated testing are subjected to failure analysis to determine the nature of the failure. From this information, improvements can be made to fix problems. With the ever increasing demands on quality and reliability, Burn in is essential to semiconductor manufacturing. With the very complex circuits of today, the semiconductor components are more sensitive to failures. The requirement of manufacturing is more stringent than it has been in the past. The demand is for more complex circuits with higher quality and reliability.

6.2 Reliability Basics

In VLSI technology, reliability is the term applied to how long a semiconductor Integrated Circuit (IC) will continue to function properly. This differs from the yield of a semiconductor part. Yield refers to the amount of parts that are manufactured correctly. The parts that pass the final functional electrical tests are the parts that are declared good and are then placed in the particular application that they were intended for. These parts are then functionally correct, as defined by the final test parameters, and should function correctly. Reliability refers to how long the particular IC will continue to function correctly. The reliability has been described as yield over time. Conversely, yield has been defined as reliability at time zero. A discussion of reliability can not take place without considering the yield. Some of the defects that affect yield also affect reliability. The same kind of defect can cause either a yield loss or a reliability failure, depending on the physical location of the defect and the severity of the defect.

To understand the reliability screening tests in VLSI technology that are in use today, an understanding of basic reliability concepts is required. The concept of a failure is key to any discussion of reliability. A failure in VLSI or ULSI IC manufacturing occurs when an IC fails to perform its desired function. This does not necessarily mean that the actual circuit itself has failed. The semiconductor part consists of a circuit in a package. The failure can occur in the package or in the interconnecting wire bonds between the circuit and the package. Even if the failure is within the circuit, this does not mean that the entire circuit is not functional. One particular part of the circuit may only be defective. Therefore, the failure in VLSI technology refers to the fact that the entire semiconductor part is not performing within specified parameters.

Another important concept in reliability is time. If time zero is arbitrarily defined to be when the part has passed the final functional test, then how is the time of failure determined? The part may not be used immediately and be put into storage before being inserted into the electronic application it was intended for. The IC may be in a circuit that is used intermittently. Therefore, the time to failure that is referred to in VLSI technology is sometimes called the mission time. This is the amount of time that the part operated within specific parameters before failure. Unless indicated otherwise, this is the time that is used in determining reliability failures.

6.3 Reliability Distributions

Failures in VLSI circuits occur at different times. One part will not fail at exactly the same time as another. If the times of failures are recorded for a population of parts over time, until all the parts have failed, a curve of the failures over time can be determined. This function of failures over time is referred to as the probability density function, f(t). An example of a normal probability density function, f(t), is shown in Figure 6.1 below.

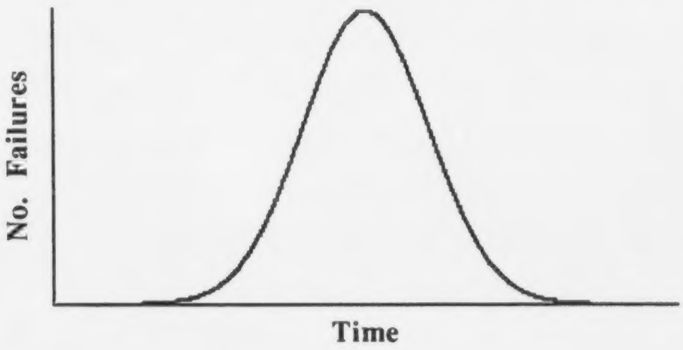


Figure 6.1 Normal probability density function, f(t)

The normal distribution example used above is an example of a short lifetime device. In most VLSI ICs, the life time of the parts is longer. The lifetimes tend to follow what traditionally has been called the "Bathtub" curve as shown in Figure 6.2.

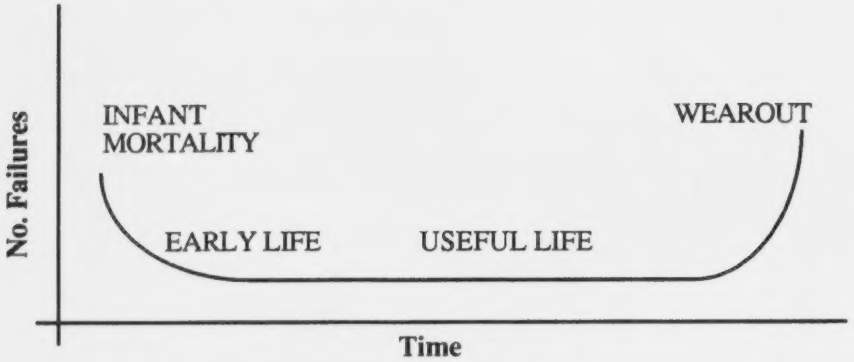


Figure 6.2 "Bathtub" failure probability density function, f(t)

The “bathtub” curve shows different regions in the lifetime of most semiconductor devices. The infant mortality region is the region where the parts have defects that are not caught by the functional final test. These parts fail early and are called the infant mortality failures. Some parts are marginal and do not fail immediately. However, they will fail early and this is an example of early life time failures. The next region is the useful life region. The failure rate remains fairly constant in this region. The number of failures increases as the parts go into the wearout region. Failures here are random and are due to random defects in the parts. The parts have survived beyond their useful lifetime and the number of failures increases drastically.

The bathtub curve is what is used in the semiconductor industry to describe the lifetime of Integrated Circuits. The useful life has a constant failure rate, $\lambda(t)$, in the useful lifetime region. In the useful lifetime region, the failure rate is low and the failures are considered to be random failures [19].

The lifetime of a number of semiconductor parts can be shown through the bathtub curve. It illustrates the number of failures over time. The bathtub curve shows that some parts fail early and some parts fail later, over a period of time. The lifetime of a single part is given by a time t. This lifetime is sometimes confused with the lifetime of a group of parts. To express the lifetime of the overall group of parts or a particular product, the median life is used, t_m or $t_{50\%}$ [20]. The median life is the time t where 50% of the population or group have failed. It is important when referring to lifetime that the lifetime of a group of semiconductor parts or product line is differentiated between the lifetime of a single part.

In addition to the probability density function, f(t), there are other statistical functions that are important in the reliability of VLSI devices. The total amount of failures that occur over time is important. The curve has a maximum of 100% or 1.0. This is because eventually all the parts in a sample will fail. The cumulative density function (cdf) denoted by F(t) is given by:

$$F(t) = \int_0^t f(x)\,dx \qquad \textbf{(6.1)}$$

Taking the example of the normal f(t) function presented earlier, the corresponding cumulative density function F(t) is shown below:

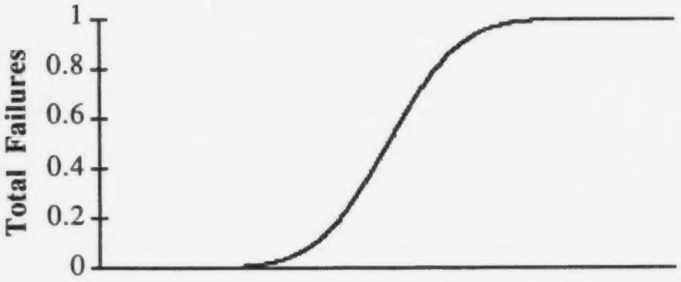


Figure 6.3 Cumulative density function, F(t) of a normal f(t) function

Figure 6.3 shows how the percentage of failed parts of a population increases with time until all the parts have failed. From the cumulative density function, the reliability probability at any given time t can be calculated. This gives the reliability of a population at time t. The Reliability function, denoted by R(t) is given by [20]:

$$R(t) = \int_t^\infty f(x)\,dx = 1 - F(t) \qquad \textbf{(6.2)}$$

By using the previous example of the normal f(t) function, the reliability function R(t) can be derived and is shown in Figure 6.4 below.

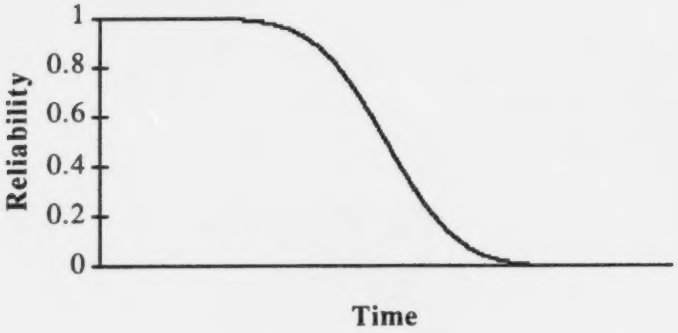


Figure 6.4 Reliability function, R(t) of a normal f(t) function

Another important concept in the reliability is the rate at which failures occur. Sometimes in reliability, the concept of "failure rate" has been misused. Failure rate sometimes has referred to the average failure rate, the instantaneous failure rate and even the cumulative failure rate. The cumulative failure rate, λ_{cum}, is defined as the total number of failures divided by the original number of devices times the total time of life [20]. Unless otherwise noted, the failure rate shall refer to the instantaneous failure rate. The development of the instantaneous failure rate is shown below:

The instantaneous failure rate, denoted by λ(t) is given by [20]:

$$\lambda(t) = \lim_{t \to 0} \frac{R(t) - R(t + \Delta t)}{\Delta t} \times \frac{1}{R(t)} \tag{6.3}$$

Solving gives [20]:

$$\lambda(t) = \frac{-dR(t)}{dt} \times \frac{1}{R(t)} \tag{6.4}$$

Rearranging equation 6.4 yields [20]:

$$f(t) = \frac{dF(t)}{dt} = - \frac{dR(t)}{dt} \tag{6.5}$$

Now, combining equations 6.5 and 6.6 yields [20]:

$$\lambda(t) = \frac{dF(t)}{R(t)} = \frac{f(t)}{R(t)} \qquad \textbf{(6.6)}$$

Solving equation 6.6 for R(t) yields [20]:

$$\lambda(t) = \frac{-dR(t)}{R(t)} \qquad \textbf{(6.7)}$$

$$\int_0^t \lambda(t)d(t) = \int_0^t \frac{-dR(t)}{R(t)}d(t) \qquad \textbf{(6.8)}$$

$$\int_0^t \lambda(t)d(t) = - \ln R(t) \qquad \textbf{(6.9)}$$

Therefore, rearranging equation 6.9 yields [20]:

$$R(t) = e^{-\int_0^t \lambda(t)d(t)} \qquad \textbf{(6.10)}$$

This equation expresses the reliability function in terms of the instantaneous failure rate. If the failure rate is constant over time, i.e λ, then the reliability function becomes [20]:

$$R(t) = e^{-\lambda t} \qquad \textbf{(6.11)}$$

This is the equation for reliability generally used in the semiconductor industry for the assumption is that the ICs are in the useful life region of the bathtub curve where the failure rate is constant. The other assumption is that other tests precipitate out infant mortality and early failures and that semiconductor ICs follow the bathtub curve.

To assume that the useful life region has a "constant" failure rate is not true. The failures in the useful life region are due to a combination of failures. The failures that result in the useful life region are a result of some infant mortality failures, some random failures that occur naturally and some freak failures due to anomalies in the complex manufacturing processes. This means that the failure rate may not be constant throughout the useful life. Therefore, there is confusion when the failure rate is described in the useful life region of the bathtub curve. Usually, failure rates quoted are not the true instantaneous failure rates but the average failure rates over the region. This average failure rate has sometimes been referred to as the "Hazard rate."

Because semiconductor parts have a relatively long life, it is usually very difficult to distinguish between the instantaneous failure rate and the average failure rate for it takes years to gather any significant amount of field data. To overcome this, most semiconductor manufacturers use some form of accelerated life testing to determine the failure rates of their products. This involves subjecting product to a higher level of stress than is normally seen and drawing some conclusions about regular product lifetime based on the results of this accelerated testing.

6.4 Failure Rate Measurements

With advances in the electronic industry and the semiconductor industry in particular, the way failure rates have been expressed has changed over time. It is important to also realize that the semiconductor parts that are referred to here are single Integrated Circuits that are usually part of a larger electronic system. Therefore, the reliability of the individual parts is usually greater than the reliability of the overall system that the parts make up. How the reliability of the system is expressed may be different from how the reliability of the individual ICs is expressed.

Before the advent of the transistor, failure rates of electron tubes were expressed in the number of failures per 1000 hours (%/1000 hours) [20]. The time referred to in hours is the actual mission time of 1000 hours. Typically, tubes had a failure rate of 5 to 10 %/1000 hours [20]. However, most semiconductor devices have a failure rate of 0.0002 %/1000 hours [20]. The unit of measurement is an awkward unit to use for semiconductor reliability lifetime. This has prompted different measurement units.

Based on the %/1000 hours is the unit of ppm, or number of failures (parts) per million device hours (#failures/1,000,000 hours). For semiconductors, this gives a lifetime measurement to the above example of 0.002 ppm. Although this is better, it still does not give easy measurement units. The semiconductor industry has been using a unit of measurement called a **FIT**. A **FIT** refers to **F**ailure un**IT**. One FIT is defined to be one failure per 10^9 device-hours[20]. Therefore, the example used above for typical semiconductor failure rates yields a failure rate of 2 FITs.

Another related unit is **RIT** or **R**emoval un**IT**. Some times when a failure occurs, a device is replaced in the system that did not need replacing. Accurate failure analysis is not performed. In semiconductor manufacturing, the reliability of a wafer manufacturing process may be the reliability that is being tested for. The failures that may be occurring may not be due to the wafer manufacturing processes, but due to other processes such as assembly processes. If adequate failure analysis is not performed, the failure rate cited will be an actual RIT rate, even though they are sometimes quoted as FIT rates.

With the increasing quality and higher demand on the reliability of semiconductor parts, the industry may again be moving beyond its measurement units. The "FIT" measurement unit may not be adequate.

6.5 Arrhenius Equation

With semiconductor reliability, one of the major assumptions is that the testing that takes place in the factory identifies all the defects that prevent the part from working initially. While there have been advances in testing and fault coverage, the testing of semiconductor parts is not 100%. When looking at reliability in semiconductors, the testing is assumed to catch all the defects that prevent the part from working initially. This is not a bad assumption, for the testing does catch almost all the defects that prevent the semiconductor parts from working.

The amount of parts that pass these tests are referred to as the yield. This separates the parts into parts that work initially from parts that have defects that prevent them from working at any time. However, even if the parts work initially, the parts are not full free of defects. It is these defects that can cause a working part not to work at some time after the initial factory testing. It is the study of these failures after factory testing that are determined to be reliability failures. The defects that cause reliability failures could be a result of contamination, improper processing or physical limitations to the construction of the semiconductors themselves. Early failures would tend to be a result of defects of the first two types, while failures in the useful life and wearout region would be a result of the physical structure of the semiconductor parts.

The question arises that how does a working part go from working in the factory to failing in the field? The part was functionally correct at the time that it left the factory but at some time later, something happened to stop the part from continuing to work. For this to happen, something happens over time. The atoms in the structure usually move or migrate over time, which causes failures. Defects in the structure of the semiconductors can enhance this process. Usually, the movement is due to some physical and/or chemical process. The rate of this process is referred to as the reaction rate. The rates that this occurs are influenced by internal stresses such as current, film stress, etc., or external stresses such as humidity, temperature and other environmental factors.

For the ion, atom or molecule can move or pass from one state or position to another, it must first overcome a potential energy barrier. The energy to overcome this barrier of height is E_A, which is in units of electron volts, eV. The probability that this transition will occur due to the body's thermal energy is proportional to [4]:

$$e^{-\frac{E_A}{kT}} \qquad \textbf{(6.12)}$$

where:

k = Boltzman's constant, 8.6×10^{-5} eV/°K
T = Absolute temperature, °K(°C + 273°)

The rate at which this reaction will occur is given by [19]:

$$R = R_0 e^{-\frac{E_A}{kT}} \qquad \textbf{(6.13)}$$

where:

R = Reaction Rate
R_0 = Constant*
E_A= Activation energy in electron-volts (eV)

k = Boltzman's constant, 8.6 x 10^{-5} eV/°K
T = Absolute temperature, °K(°C + 273°)

The above equation is known as the Arrhenius equation [4]. The term R_0 is listed in the equation as a constant. However, this is just an approximation. R_0 can also be a function of temperature. It has generally been determined that the temperature dependence of R_0 is small when compared to the exponential term in equation 6.13 [4]. This assumption that R_0 is constant is used throughout the semiconductor industry. This equation adequately describes the rate of processes that are responsible for the failure of many semiconductor devices such as ion drift, impurity diffusion, intermetallic compound formation, molecular changes in insulating materials, etc.

6.6 Eyring Equation

The Arrhenius equation relates the reaction rate to the environmental temperature. It is useful in determining the effects of temperature stress to lifetime. However, temperature stressing is not the only type of stressing used in accelerated semiconductor testing. Other stresses have been used to accelerate the time to failure of semiconductor ICs.

Another model that has been proposed for accelerated testing is the Eyring equation. This equation is given below in terms of the component hazard rate:

$$\lambda = A\left(e^{\frac{B}{kT}}\right)\left(e^{S\left(C+\frac{D}{kT}\right)}\right) \tag{6.14}$$

The above equation actually consists of three parts. A, B, C and D are constants, T is temperature in degrees Kelvin, and k is Boltzman's constant. The first term, $e^{\frac{B}{kT}}$, is the Arrhenius term for temperature dependence and comes from the Arrhenius equation. The second term is e^{SC} which is the stress term for the stress other than temperature. The final term in the equation is $e^{S\frac{D}{kT}}$, which is the term in the model that refers to the interaction of the stress and the temperature effects. This is one model that has been proposed to describe the interaction of additional stresses, these stresses are used in accelerated testing of semiconductor devices.

Some of the stresses that have been used for accelerated testing are humidity, voltage, and current. These stresses have been used to accelerate the failure rate of semiconductors. Specifically, a model that has been proposed for the effect of voltage has been modelled as [22]:

$$R(T,V) = R_0(T)V^{\gamma(T)} \qquad \textbf{(6.15)}$$

Where the above equation contains a temperature acceleration term and a voltage acceleration term, a similar relationship has been proposed for current stresses:

$$R(T,V) = R_0(T)J^{\gamma(T)} \qquad \textbf{(6.16)}$$

In both equations 6.15 and 6.16, the term, $R_0(T)$, is usually the Arrhenius equation. The other two terms, $V^{\gamma(T)}$ and $J^{\gamma(T)}$, describe the voltage and current stress factor as some power of the voltage, V, and the current density, J. The power term, γ is usually some value between 1 and 4.5 [22].

An example of an accelerated current stress is the relationship for electromigration. It has the current acceleration factor in the equation and an Arrhenius temperature acceleration factor.This equation is again repeated below [11]:

$$MTTF = A*J^{-n}e^{(Q/kT)} \qquad \textbf{(6.17)}$$

6.7 Accelerated Reliability Testing

The Arrhenius equation shows the relationship between the reaction rate and the effect of temperature on reaction rate. It is this relationship that is used in the semiconductor industry to accelerate lifetime tests. This is best illustrated by the following comparison lifetime graph of two parts at two different temperatures during their useful lifetimes as shown in Figure 6.5.

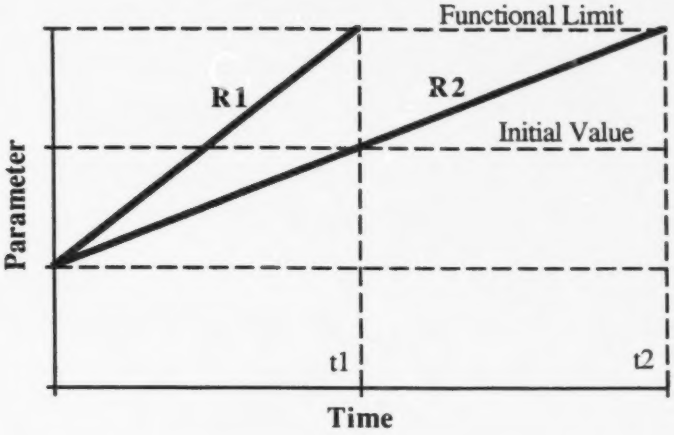


Figure 6.5 Component lifetime at two different reaction rates, R_1 & R_2 [19]

The first part started at the Initial value of the measurement parameter until at time t_1, the functional limit of the value is exceeded and the part fails. The parameter could be the resistance of the interdielectric vias. With the mechanism of electromigration, the metal in the vias can move over time and the resistance will increase until the part fails. This is at the functional limit of resistance. The temperature that this occurs at is T_1 at the reaction rate R_1.

Consequently, the second part is in an environment at temperature T_2 and a reaction rate of R_2. The part lasts for time t_2. The change in the parameter is the same in both cases but the second part lasts longer for the temperature T_2 is lower than T_1, and the rate R_2 is lower that R_1. It must be noted that the assumption in the above graph is that the reaction is linear in time.

In the above example of the via resistance, the electromigration rate will occur at a slower rate at the lower temperature. Since the change in the parameter is the same, the following is true [19]:

$$R_1t_1 = R_2t_2 \tag{6.18}$$

and if the functional limit is where the part fails, then the time to failure can be defined as t_f. Then equation 6.18 can be written as [19]:

$$Rt_f = C \tag{6.19}$$

Where C is the constant that corresponds to the parameter that causes the part to fail its functional limit. From this [19]:

$$t_f = \frac{C}{R} \tag{6.20}$$

And then from equation 6.13 [19]:

$$t_f = Ce^{\frac{E_A}{kT}} \tag{6.21}$$

$$\ln(t_f) = C + \frac{E_A}{kT} \tag{6.21}$$

By using the above example of two different reaction rates, then equation 6.22 becomes the following [19]:

$$\ln(t_1) = C + \left(\frac{E_A}{k}\right)\left(\frac{1}{T_1}\right) \tag{6.23}$$

$$\ln(t_2) = C + \left(\frac{E_A}{k}\right)\left(\frac{1}{T_2}\right) \tag{6.24}$$

Subtracting equation 6.24 from equation 6.23 yields [19]:

$$\ln(t_1) - \ln(t_2) = \left(\frac{E_A}{k}\right)\left(\frac{1}{T_1} - \frac{1}{T_2}\right) \tag{6.25}$$

or [19]:

$$\frac{t_1}{t_2} = e^{\left(\frac{E_A}{k}\right)\left(\frac{1}{T_1} - \frac{1}{T_2}\right)} \tag{6.26}$$

This relationship allows the comparison of two different reaction rates. This is the basis of all accelerated temperature testing used in semiconductors today. The times t_1 and t_2 correspond to the times of an unaccelerated lifetime of the product (t_1) and the accelerated lifetime of the product (t_2). The relationship of t_1 and t_2 is also refereed to as the acceleration factor.

$$\text{Acceleration} = \frac{t_2}{t_1} \tag{6.27}$$

The above development assumed that the reaction was linear in time. But not all the failure mechanisms that are seen in semiconductor manufacturing are linear in time. These failure mechanisms can be represented by a t^n function. Substituting this in to equation 6.19 yields [19]:

$$R t_f^n = C \tag{6.28}$$

$$t_f = \left(C e^{\frac{E_A}{kT}}\right)^{\frac{1}{n}} \tag{6.29}$$

$$\ln (t_f) = C + \frac{E_A}{nkT} \tag{6.30}$$

And going through a similar derivation as shown above:

$$\frac{t_1}{t_2} = e^{\left(\frac{E_A}{nk}\right)\left(\frac{1}{T_1} - \frac{1}{T_2}\right)} \tag{6.31}$$

Equation 6.31 shows the acceleration for failure mechanisms that are not linear in time.

It is the principle of applying stress to accelerate lifetime testing that is the basis of almost all semiconductor reliability testing. With the lifetimes in tens of years, it is impossible to gather data about the reliability of a product that is made today. Through the application of higher than normal stress, the reaction rate of the failure mechanism is accelerated. This allows reliability testing to be done within a reasonable time.

If a company makes a particular semiconductor product, a microprocessor for example, that company and its customers want some assurance that the product will work reliably over time. By subjecting a sample of the microprocessors to some accelerated testing, information about the median lifetime of the microprocessors can be obtained by the use of the acceleration rate equations shown above. Table 6.1 shows some of the acceleration factors that are used in semiconductor accelerated lifetime testing and their approximate activation energies.

Table 6.1 Acceleration factors for semiconductor reliability [20]

Device Association	**Process**	**Relevant Factors**	**Accelerating Factors**	**Acceleration (E_A =Activation Energy)**
SiO2 and Silicon - SiO2 Interface	Surface Charge Accumulation	Mobile Ions, V, T, Qm	T	1.0 eV
	Dielectric Breakdown (TDDB)	E, T	E(T)	0.35 eV
	Charge	E, T, Qf	E, T	1.3 eV
	Hot Carrier Trapping	E, T, Qot	E, T	-0.06 eV
Metallization	Electromigration of Al	T, J, A Gradients of T and J, Grain Size	T, J	0.5 eV J^n ; n = 2±10%
	Electromigration of Si in Al	T, J, A	T, J	0.9 eV
	Corrosion	Contamination, Humidity (H) V, T	H, V, T	Strong H Effect 0.8 eV
	Contact Degredation	T, Metals, Impurities	Varied	1.8 eV
Bonds and Other Mechanical Interfaces	Intermetallic Growth	T, Impurities, Bond Strength	T	Al-Au: 1.0 eV
	Fatigue	Temperature Cycling, Bond Strength	T Extremes in Cycling	Al-Au: 1.0 eV
Hermeticity	Seal Leaks	Pressure Differential, Temp. Cycling	Pressure ΔT	Al-Au: 1.0 eV

LEGEND	H=Humidity	T=Temperature	V=Voltage	A=Area
	E=Electric Field	J=Current Densisty	Q=Oxide Charge	

6.8 Semiconductor Reliability Testing

By use of accelerated testing, Semiconductor manufacturers use a sampling technique to assure the reliability of the product produced and the processes used. The nature of the sampling varies from manufacturer to manufacturer, depending upon the reliability results desired and the level of acceptable defects. Most manufacturers do reliability testing as a method of ensuring that the ongoing processes and designs meet minimum standards. Any time new technologies in manufacturing are introduced or changed in a major way, a number of parts are produced specifically for reliability testing before any parts are manufactured for customer use. If the parts meet the reliability standards, the process or design is then considered qualified. The process of manufacturing a specific lot to test out changes or new technology is called "qualification".

The "qualification" lots are processed through the proposed manufacturing flow and assembled into packages. These parts are then subjected to accelerated life time testing. The corresponding lifetimes of the product are calculated with the use of the Arrhenius equation. If the results are acceptable, then the process is considered qualified. Qualification is generally done when one or more of the following changes occur [19]:

1. New processes or technology
2. Modifications to existing processes
3. Design revision
4. Wafer fab Process/product transfers or technology transfers
5. New part packaging or packaging technology/process

Qualification involves stressing the final semiconductor product in a package. This is done through a number of accelerated lifetime tests, designed to precipitate out the failure modes in semiconductors. An example of a set of reliability tests done by a semiconductor manufacturer is shown below [19]:

1. High Temperature Dynamic Lifetest
2. High Voltage Dynamic Lifetest
3. Low Temperature Dynamic Lifetest
4. High Temperature Storage
5. High Temperature/Humidity Lifetest
6. Temperature Cycling
7. Vibration/Shock Testing
8. Soft Error/Radiation Hardness Testing

The tests listed above are an example of what a particular process or technology may be subjected to for qualification purposes. Not all of the tests may be performed on all qualifications within a single semiconductor manufacturer. The nature of the processes, designs and technology determines what accelerated lifetime testing will be performed. A description of these tests follows.

6.8.1 High Temperature Dynamic Lifetest

This test involves stressing a sample of semiconductor parts through the use of temperature and heat. Parts are powered and inputs are functionally set to provide current stresses. This occurs on circuit boards placed in an oven, usually at 125 °C. The parts are subjected to this condition for 1000 up to 2000 hours total. During the testing period, all the parts are removed from the burn in oven and functionally tested. This occurs at various set times through out the testing period. The time to do this functional testing is not included in the 1000 or 2000 hours of lifetime testing. The period that this testing occurs tends to be of a logarithmic nature. This is because the reaction rate from the Arrhenius equation is of a logarithmic nature. An example of the testing period would be after 48 hours, 168 hours, 500 hours, and 1000 hours for a 1000 hour burn in period. This period of testing is similar for all different stresses based on the Arrhenius relationship.

The High Temperature Dynamic burn in is referred to by different names through out the semiconductor industry. Another manufacturer refers to the test as High Temperature Dynamic Operation Life, for example. This test is what the industry uses to evaluate the field lifetime of semiconductor parts. The test is an industry standard and is listed in MIL-STD-883C. The semiconductor industry has conformed to these military standards and these have become the standards for the industry. Most of the accelerated stress test listed here meet the requirements of MIL-STD-883C.

The dynamic burn in accelerated life test is the industry standard for determining the field lifetimes of semiconductor product. The resulting data is used to determine infant mortality failures, early lifetime failures, random failures and wear-out failures. Failure analysis and empirical testing are used to determine the failure mechanisms and the activation energy of the observed failures, using the Arrhenius equation. With the nature of the failure know and the corresponding activation energy, the lifetime of the product can, therefore, be calculated, using the acceleration factor equation, equation 6.27.

6.8.2 High Voltage Dynamic Lifetest

This test involves stressing the semiconductor parts for 1000 to 2000 hours with a higher than normal operating voltage, typically 7 to 8 volts. This test tends to verify the failure mechanism and rates of the long term reliability failures. This test is also useful in identifying the acceleration factors. The test conditions are the same as the high temperature dynamic lifetest, with the exception being the higher voltage used.

6.8.3 Low Temperature Dynamic Lifetest

This test is designed to precipitate a particular failure mechanism and is not done by all semiconductor manufacturers, depending upon their technology and reliability requirements. The low temperature test is designed to show failures in semiconductor parts, particularly CMOS parts, due to hot carrier/ hot electron injection. As shown in Table 6.1, the activation energy for this failure is negative. This means to accelerate the effect of this failure mechanism, lower temperatures are required. This test is usually done for a 1000 to 2000 hours at a temperature of -10℃, with a high voltage (7-8 volts) and a high duty cycle. The duty cycle refers to the amount of switching that occurs at the inputs of the parts. The inputs are switched between a logic 0 and a logic 1 (7-8 volts) quickly and repeatedly to accelerate the failure.

6.8.4 High Temperature Storage

This test uses only temperature for stressing. The parts are subjected to high temperatures with no applied voltage bias. Typical values are 160°C for 1000 hours for parts in plastic packages and 250°C for 500 hours for parts in ceramic (hermetic) packages. This test tends to show package related failures such as bond integrity and package mechanical instability. In addition is also shows some wear out failure mechanisms related to the wafer processes such as ionic contamination.

6.8.5 High Temperature/Humidity Lifetest

The purpose of this test is to determine the moisture resistance of the semiconductor part. This is done on plastic-encapsulated parts because they are susceptible to failure due to moisture. Some of the effects that are tested for are [20]:

1. Mobile ions on surface devices
2. Chemical corrosion of the metallization
3. Electrolytic corrosion with applied bias
4. Electrolytic corrosion with dissimilar metals

The components are placed in a chamber usually at 85°C with an 81% relative humidity. The parts are biased with a nominal voltage of either 5 volts or 0 volts. There is minimal power applied and what little bias is applied is used to enhance the electrolytic corrosion that can occur in the presence of moisture. This maximizes the effect of corrosion due to electrolytic effects.

6.8.6 Temperature Cycling

This test involves cycling the temperature of the components from a very cold environment to a very hot environment. The purpose of this test is usually to detect various mechanical instabilities in the assembly bonds and die attach as well as stress related issues in the die itself such as microcracks. There is no other stresses applied to the part. The parts are typically cycled from -55°C to 150°C or from 0°C to 125°C. The number of cycles used are from 500 to 1000 cycles. This is usually done in a dual chamber system where one chamber is cooled to the desired cold temperature and the other chamber is heated. The parts are physically transferred back and forth between the two chambers with the use of robotic transfer systems.

6.8.7 Vibration/Shock Testing

For some qualifications and applications, a vibration and shock stress testing is used to determine if the parts can withstand vibration and shock. Semiconductor parts subjected to these tests are parts designed for applications where there is a great amount of physical vibration and shock. Unfortunately, at this time, no models exist for accelerated stress testing in this area. This means that the testing tends to be more of the infant mortality data gathering and not of the useful lifetime determination.

This testing is used in applications where there are environments that produce high levels of vibration and/or frequent shock. Some of these application environments may be:

1. Read/write electronic circuits assembled on the head of high density, high speed magnetic disk equipment
2. Control IC circuits on jet engines
3. Control IC circuits on rockets, missiles and spacecraft.

6.8.8 Soft Error/Radiation Hardness Testing

Another reliability test that is only done on components destined for a specific application is the soft error or radiation hardness testing. To test for the effect of radiation, parts are exposed to a highly radioactive source. This is only done to parts that will require exposure to high levels of radiation because of the difficulty of working with radioactive sources.

The sensitivity of the particular product to radiation can vary with the process and the applications. Processes are characterized to see the effect on the radiation susceptibility. Since the part's operating voltage conditions can affect the radiation sensitivity, the operating conditions of the part are characterized [19]. The types of applications that would require this type of characterization are the aerospace industry, satellites and the military.

6.9 Reliability Monitoring

In addition to the qualification tests, the use of on-going reliability monitoring is important to ensure the on-going reliability of the product. Even though principles of Statistical Process Control (SPC) are used in semiconductor manufacturing today, the identification of defects is important. Reliability monitoring will identify the particular failure mechanisms that will cause the product to fail first. Improvements and developments can then continue to lessen the effect of the failure mechanism in question, and improve the reliability and quality of the product.

The second function that reliability monitoring performs is the assurance that there are not reliability problems in the existing manufacturing process. Although the best way is to build quality into the product through the application of SPC and design of experiments, traditional reliability monitoring is still important to determine the failure modes in todays semiconductor components. This monitoring insures that the processes used in the manufacturing of the device are in control and that there are no unknown factors affecting the product reliability.

Most semiconductor operations conduct High Temperature Dynamic Lifetests as the main test for reliability monitoring. This accelerated test will tend to precipitate out most of the failure mechanisms. The results of these tests will also be used to calculate the FIT rate of the processes and product being tested. Because the parts are stressed, the components used in the testing are never put into the applications that they were designed for. The chips are a sample from actual product that is being produced in the manufacturing line. Other accelerated tests may also be used for reliability monitoring. It depends on the manufacturer in question and the requirements of the customer whom the components are being manufactured for. The accelerated testing is only part of an overall reliability monitoring program.

Wafer Level Reliability (WLR) tests are used to monitor reliability of the product. These tests are usually done on specific test structures that have been designed for a particular failure mechanism. The tests are done usually through a very highly accelerated voltage or current stress on the test structure. The testing is done with the use of parametric testers, and the testing time is in the order of minutes to even in some cases, days. The data that is derived from Wafer Level Reliability testing is not as reliable as the data from the High Temperature Dynamic Lifetest. The Dynamic Lifetest will tend to precipitate a number of different failure mechanisms where the Wafer Level Reliability testing only tests for a particular failure mechanism. The advantage is that WLR testing only takes a few minutes where Dynamic Lifetest takes 1000 hours or more. More monitoring can be done with Wafer Level Reliability testing, yielding more data and information about the particular failure mechanisms being testing for.

The other advantage of Wafer Level Reliability is that Wafer Level Reliability tests are done on separate test structures that are heated through their own design and test programs. The test structures are self-heating, and only the test structures are subjected to stressing. With Wafer Level Reliability, no product is destroyed in accelerated testing. Tests such as the Dynamic Lifetime test destroy product in the test.

The Dynamic Lifetest and the other traditional burn in tests listed previously are useful in identifying the predominant failure mechanisms that should be monitored for. The burn in tests will help identify the failure mechanisms that occur the most often and that could cause the failures of components first. From this information, tests and test structures can be developed to monitor for these failure mechanisms through Wafer Level Reliability. Even more important, improvements in the manufacturing processes, design and the technology itself can occur to reduce the effect on the reliability of the product of these failure mechanisms.

Table 6.2 below shows actual results of a reliability program as quoted by Intel. This table shows the life time for a number of technologies. The lifetime refers to the time at which 50% of the population would have failed. For example, in the 125℃ Dynamic burn in, the Device Hours is quoted to be 4.29×10^7 hours. Now the test is only for 1000 or 2000 hours. As shown before, the failure rate, l, has been assumed to be constant in the useful life time, even for the accelerated reaction rate. By knowing the life time (e.g. 1000 hours) and the number of parts that has failed in the test, the failure rate can then be calculated. Then this accelerated l is used to calculate the Device Hours,which are the amount of time it would take for 50% of the parts to fail at 125℃, or the t_m at 125℃.

From the example previous, at 1000 hours, there would be only 0.23% of the parts failing. This means that for every 10,000 parts put through burn in, only 23 failures would be allowed to maintain the quoted failure rate. Through the use of Arrhenius equations the t_m at 125℃ can then be used to calculate the medium life time, t_m at the desired specified temperature, 55℃ in this case. From the specified lifetime, the failure rate can, therefore, be calculated for a specific failure with a corresponding activation energy. The results of a reliability program are listed below in Table 6.2.

Table 6.2 Lifetime data [19]

Tech.	No. Lots	Device Hours	Equivalent Device Hours at 55°C, Ea=0.3, 1.0 eV		Failure Rate at 55°C, Ea=0.3, 1.0 eV	
		at 125 °C	for 0.3 eV	for 1.0 eV	for 0.3 eV	for 1.0 eV
A	409	4.29E+07	2.37E+08	1.35E+10	70	0
B	299	4.77E+07	2.50E+08	1.21E+10	110	0
C	12	2.33E+06	1.23E+07	5.92E+08	80	0
D	102	2.72E+07	1.76E+08	1.36E+10	50	0
E	155	1.31E+07	5.98E+07	8.70E+08	190	0
F	167	7.24E+07	4.98E+08	3.10E+10	160	0
G	213	2.01E+07	1.08E+08	1.29E+09	50	0
H	261	4.06E+07	2.14E+08	9.36E+09	40	0

Process monitors within the manufacturing processes are important for a semiconductor reliability monitoring program. By monitoring the quality of the metal deposited, the incoming materials, the manufacturing processes, packages, reticles, etc.; the quality and the reliability of the product can be maintained. This is why most semiconductor manufacturers have quality and reliability personnel and departments. Most manufacturers today work on the systems used in manufacturing. The aim is to build in quality and reliability, not inspect it in. By controlling and concentrating on the inputs to the manufacturing processes, the quality and the reliability of the outgoing product is assured. Burn in, therefore, acts as more the monitor of this reliability program and is used to identify failure mechanisms that should be monitored and improved.

Infant mortality burn in has also been a part of a traditional reliability monitoring program. Here parts are subjected to a bake under no biasing condition. The temperatures and time of the bakes vary, but the intent of this burn in bake is to precipitate any infant mortality failures, and move the product into the useful life region of the bathtub curve.

With the improved processing technologies, the reduction in design line widths, and the increased complexity of VLSI and ULSI components, there is now fewer infant mortality failures. Testing procedures have also improved. More failures are precipitated in the initial product testing, and the remaining defects tend to be of the type that will cause failure during the useful life of the product. Therefore, the infant mortality burn in does not yield as much data or cause the same amount of failures that have been seen in the past. Depending on the product and the technology, this burn in may not be even used any more.

6.10 Summary

Burn in of semiconductor product is an essential part of an overall reliability and quality program for a manufacturer. By the use of burn in, the failure mechanisms can be accelerated such that they will be identified in a reasonable amount of time. Burn in allows potential failures that could occur in the field to be identified long before they would show up in the applications. With the lifetime of semiconductor components being a number of years, accelerated testing through the use of burn in is one method of identifying the particular failure mechanisms in a timely fashion.

Quality of the final product is important in the semiconductor manufacturing industry today. There has been a shift away from the traditional burn in and towards the building in of reliability. Burn in remains an important part of an overall reliability and quality program. It is still used to calculate the expected field lifetimes through the use of accelerated testing models such as Arrhenius Equation and Eyring equation. From these models, the acceleration factors are calculated and the corresponding lifetimes and failure rates.

The use of traditional burn in techniques remains important in the semiconductor industry. With the development of Wafer Level Reliability tests, it has been suggested that burn in would eventually be replaced by WLR testing. This is not so. Burn in accelerates failures at a lower rate than does WLR testing. Therefore, the lifetime data tends to be more accurate than the WLR testing. Also, burn in methods tend to identify failure mechanisms and defects that affect reliability. Wafer Level Reliability only tests a particular defect that the particular test and corresponding test structure is designed to.

Wafer Level Reliability will not replace burn in for lifetime testing and will be an important supplement to the overall reliability monitoring program. WLR testing supplies a lot more data quickly than burn in about the reliability of certain failure mechanisms. It is developing into a useful tool for insuring the reliability of product. Burn in lifetime testing will continue to be an important part of semiconductor manufacturing.

CHAPTER VII

EVALUATION OF WAFER LEVEL RELIABILITY TEST DATA

7.1 Introduction

In understanding the application of Wafer Level Reliability to a semiconductor manufacturing environment, the knowledge of the limitations and use of Wafer Level Reliability is required. To investigate this, a series of manufactured runs through a standard processes were commissioned. The goal of these runs was to determine if Wafer Level Reliability testing could show specific failure mechanisms.

Long term reliability burn in stress tests such as High Temperature Dynamic Lifetest (HTDL), High Voltage Dynamic Lifetest (HVDL), Low Temperature Dynamic Lifetest (LTDL), High Temperature Storage (HTS), High Temperature/Humidity Lifetest (HTHL), Temperature Cycling (TC), Vibration/Shock Testing (VST) and Soft Error/Radiation Hardness Testing (SERHT), are used to determine the major reliability failure modes in product. From these failure modes, test structures can be designed and tested at wafer level to determine relationships between WLR and Lifetime data. These WLR tests can include JRAMP, VRAMP, BVOX, QBD, SWEAT, BEM, and hot electron testing.

In the gate oxide reliability testing presented in this chapter, the gate oxide was evaluated through the use of WLR tests - VRAMP, QBD and BVOX - and through the lifetime burn in data - High Temperature Dynamic Lifetime, High Temperature Storage, and High Temperature/Humidity Lifetest.

7.2 Experimental Procedure

The three areas of ASIC gate array manufacturing was investigated. These are: Foundry manufacturing, Metallization manufacturing and Assembly manufacturing. The terms Foundry, Metallization and Assembly apply to the LSI Logic process flows which were used to produce the test wafers. The basic process flow of these three areas are described in the following sections.

7.2.1 Test Run Processing

The first processes used to manufacture the wafers are the Foundry processes. The

Foundry manufacturing processes take the bare silicon wafers and defines the transistors on them. This involves the various implants to define the p and n junctions, the gate oxidation, to define the polysilicon structures, and the final protective oxide over top of the transistors. After the Foundry processing, the wafers have fully functional transistors on them protected by a silicon dioxide layer doped with both boron and phosphorus. This protective layer is called Boron Phosphorus Silicate Glass (BPSG).

The next series of processes in the production of the test wafers, the Metallization process, start by taking Foundry wafers out of inventory. Contacts through the protective BPSG to the underlying transistors are first formed. The first layer of metal interconnect is then deposited, masked and etched. The interlayer dielectrics is then deposited and planarized to flatten out the topology of the wafers. The vias are then formed through this interlayer dielectric. The second metal interconnect is then deposited, masked and etched. The final protective layer called the topside passivation is then deposited. The final pad mask layer is then masked and etched. The circuitry on the die is now complete. The metallization process that the test wafers were manufactured with is a two layer metallization process. The wafers then go through a parametric test to determine if basic transistor parameters are functioning. After this test, the test wafers were submitted to a wafer functional test to determine functioning die. This completes the metallization processes.

The third set of processes is the Assembly process. Here, the wafers are cut into the individual die that are then placed into packages. The die from cut wafers are placed into individual packages to be wire bonded. Wire bonding is the process of connecting the package to the die. Small wires are attached through the use of ultrasonic welding. These small wires attach the bonding pads on the die with the pads on the package. These package pads are connected internally in the packages to the leads on the outside of the package. The package is then sealed and marked with the correct identifying marking. The completed part is then tested again functionally at the Final Test operation to insure that the part is still functionally operational.

For the Wafer Level Reliability test runs, the Foundry process flow that was used is as follows:

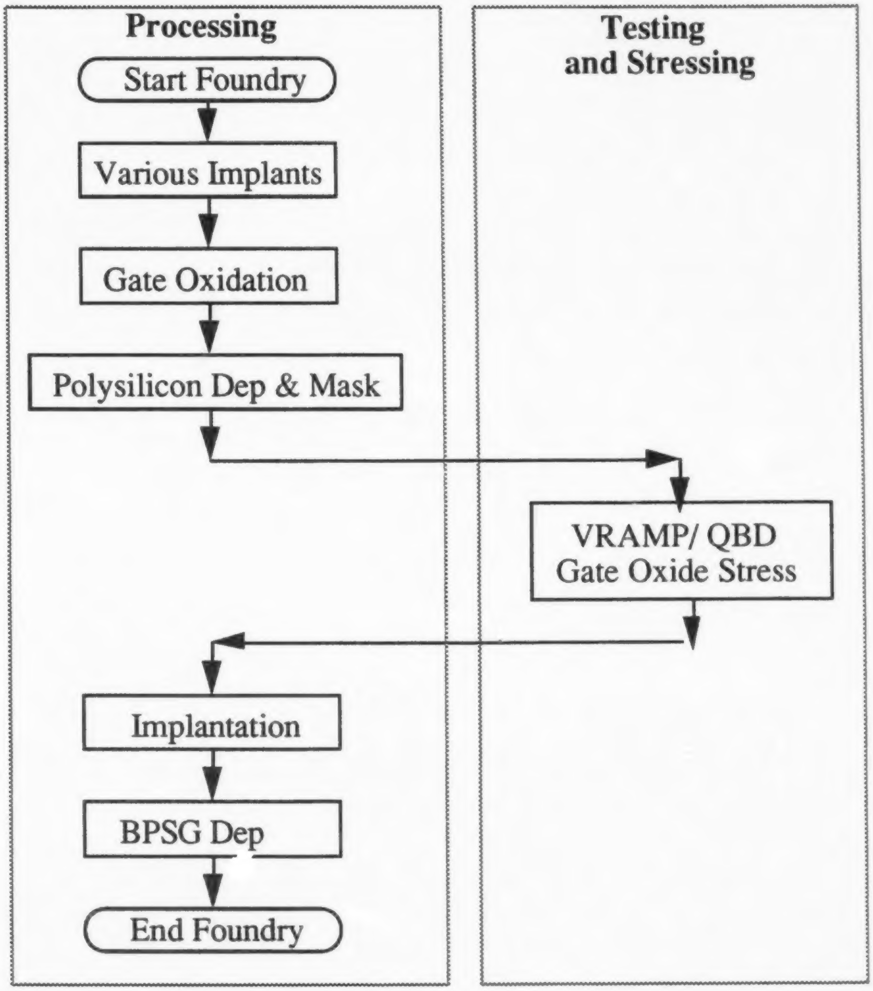

Processing
Testing and Stressing
Start Foundry
Various Implants
Gate Oxidation
Polysilicon Dep & Mask
VRAMP/ QBD Gate Oxide Stress
Implantation
BPSG Dep
End Foundry

Figure 7.1 Test run foundry process flow

The test foundry runs went through a standard manufacturing flow except for additional testing and stressing after the polysilicon masking steps. Here the gate of the transistors has been formed and this is the first place in the manufacturing processes that the parts can be stressed.

The Metallization process flow that was used is as follows:

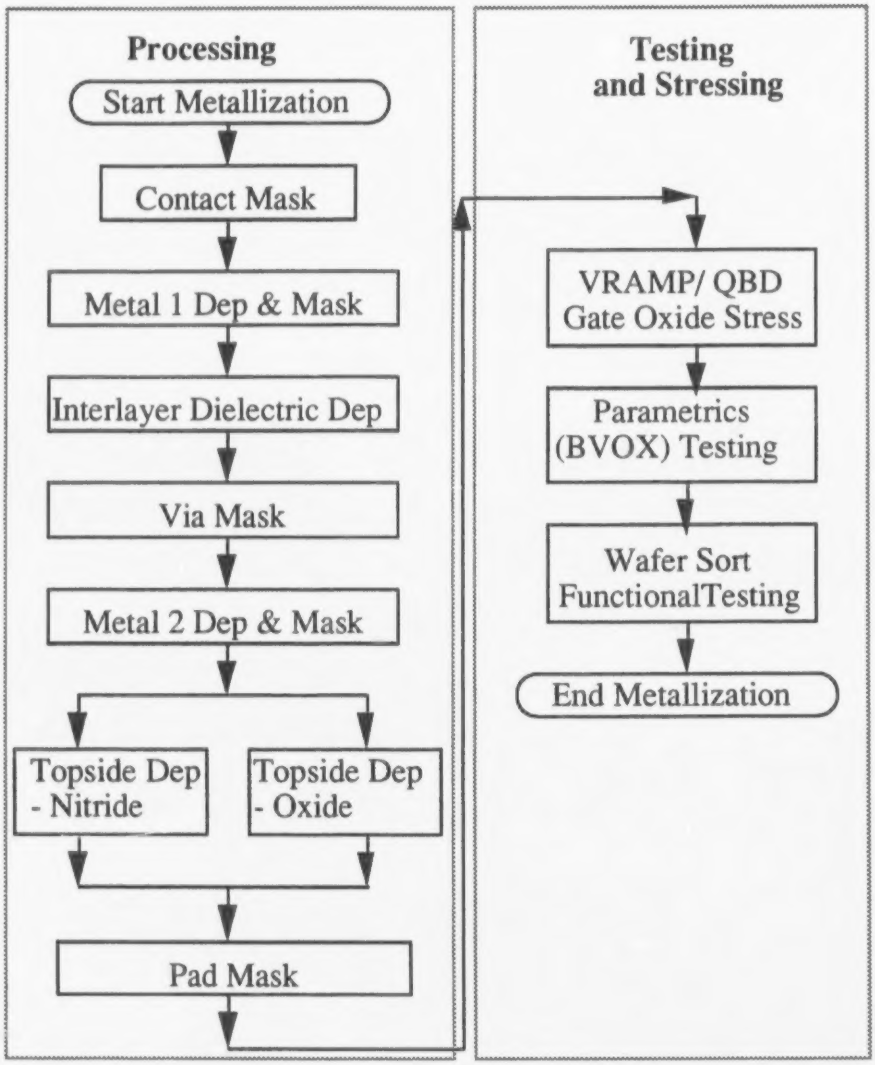

Processing
Start Metallization
Contact Mask
Metal 1 Dep & Mask
Interlayer Dielectric Dep
Via Mask
Metal 2 Dep & Mask
Topside Dep - Nitride
Topside Dep - Oxide
Pad Mask
Testing and Stressing
VRAMP/ QBD Gate Oxide Stress
Parametrics (BVOX) Testing
Wafer Sort FunctionalTesting
End Metallization

Figure 7.2 Test run metallization process flow

Through the Metallization process flow, the test runs went through standard processing. The runs that were destined to have ceramic packages for reliability stressing received silicon oxide as a passivation material. The runs that were to be packaged in plastic packages for reliability stressing received silicon nitride as a passivation material. This is the standard process flow. The silicon nitride is more resistant to moisture and is used in plastic packaging applications. The test structures on the test part were stressed through the use of VRAMP, BVOX, and QBD testing before the die were sorted for functional operation. The Wafer Level Reliability test parts went through the standard assembly process as shown in Figure 7.3.

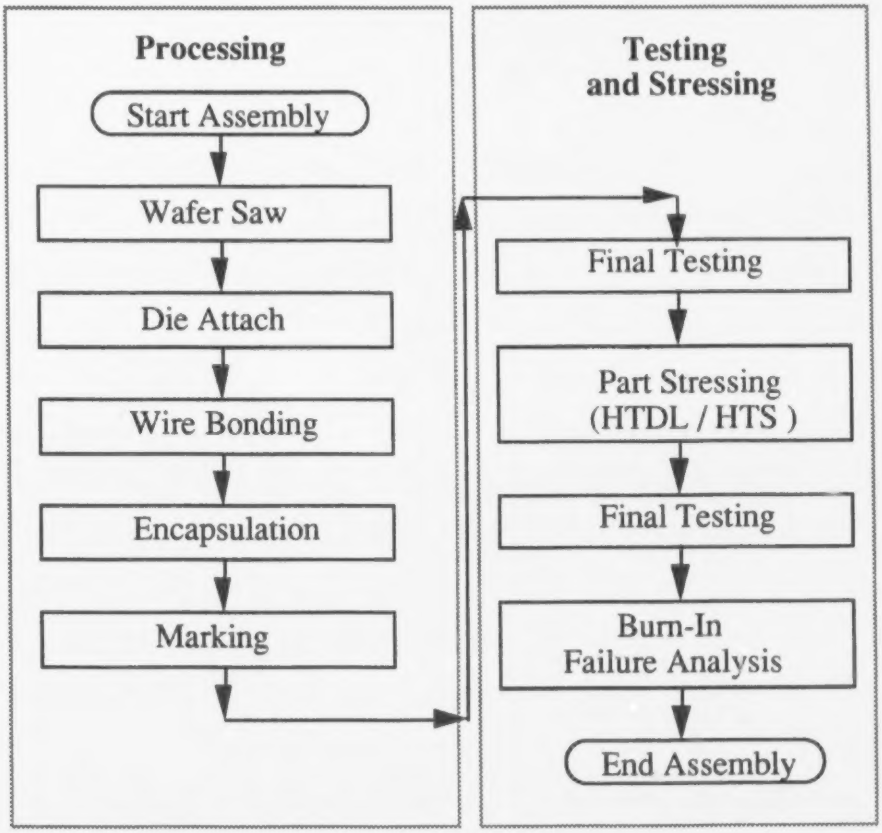

Processing
Start Assembly
Wafer Saw
Die Attach
Wire Bonding
Encapsulation
Marking
Testing and Stressing
Final Testing
Part Stressing (HTDL / HTS)
Final Testing
Burn-In Failure Analysis
End Assembly

Figure 7.3 Test run assembly process flow

The test runs were assembled into both ceramic and plastic packages. After assembly, the parts went through the final test functional test to insure that the assembly process had been successful. Then the parts were subjected to the traditional reliability stressing (burn in) of either HTDL (High Temperature Dynamic Lifetest) or HTS (High Temperature Stress). The parts were again tested throughout the stressing procedure to determine failures. The failures were then analyzed to determine the exact nature of the failures.

7.2.2 Test Wafer Layouts

To fully test the relationship between the Wafer Level Reliability structures and the operation of standard ASIC gate array product, a special reticle set has been made with the Wafer Level Reliability test structures on one die location and a monitor logic chip on the adjacent die location. An example diagram of a typical wafer layout is shown in Figure 7.4

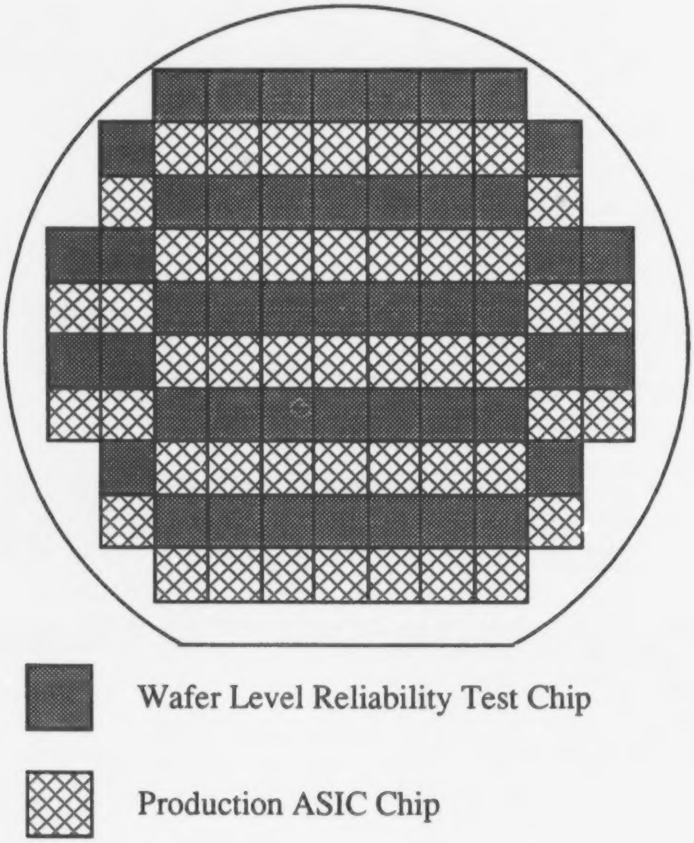


Figure 7.4 Test wafer layout

The production ASIC test chip contains logic circuitry and memory for process evaluation. This chip has functional logic circuits on it and can be tested at wafer sort. During the test runs, the WLR structures on the Wafer Level Reliability test chip are stressed to obtain data. The production monitor chip is the chip that is tested functionally at wafer sort. This die is then assembled into either plastic and ceramic parts and stressed by either the High Temperature Dynamic Lifetest (HTDL) test or the High Temperature Storage (HTS) test.

Through the use of the separate test chip, a better understanding of the structures will develop. But the use of such a large amount of wafer area for a dedicated test chip is too expensive for regular ongoing production. Between every die is an area that is used for elements that are not necessary to the function of the circuit, but are necessary to the construction of the circuit. This area between the die is known as the scribe line. Here, the alignment targets for photolithography are defined, the basic sample transistors for parametric testing, yield testing structures and some Wafer Level Reliability structures. From the analysis of Wafer Level Reliability test results from the test chips and the analysis of failure modes from burn in of adjacent product ASIC chips, the prevailing failure modes can be determined. Associated structures with these failures can be then chosen and put alongside every die in the scribe line. An example layout is of a wafer with a large scribeline with WLR structures in the scribe line is shown in Figure 7.5

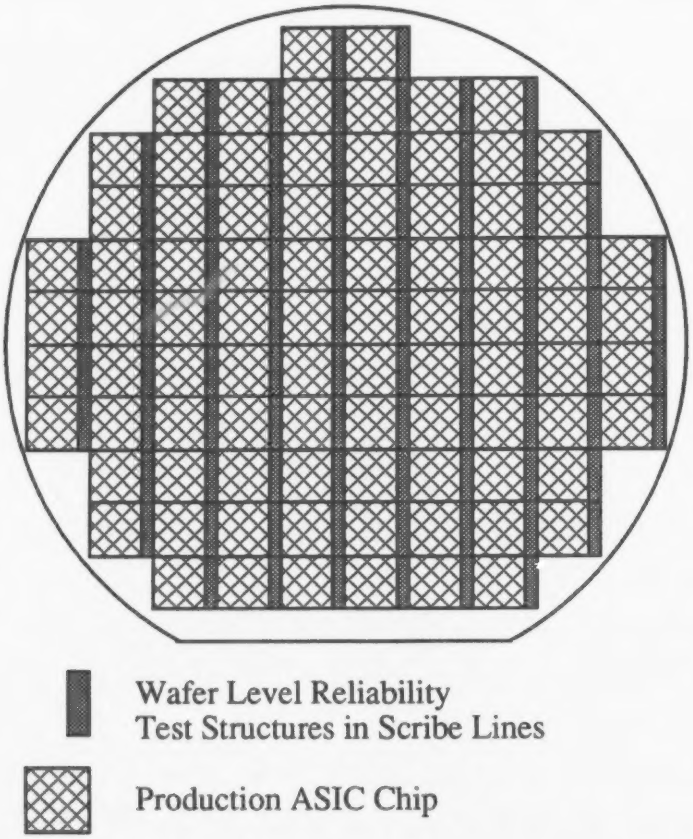


Figure 7.5 Example of Wafer Level Reliability scribe line test structures

The layout above shows a large scribe line structure adjacent to production ASIC die. The Wafer Level Reliability structures only make up a portion of the structures found in this scribe line. The rest of the scribe line structures would include yield structures, alignment structures and parametric test structures. In the test runs that were done to determine the gate oxide reliability, some existing structures in the scribe lines were used to determine the gate oxide Wafer Level Reliability data.

7.3 Empirical Testing Results

The runs that were used in this evaluation of Wafer Level Reliability came from five foundry lots. From these five foundry lots, six metallization runs were made. The runs were selected to have either oxide or nitride passivation, depending upon which package option was used. This is as per standard production procedure. The test results are shown in Table 7.1

Table 7.1 Wafer Level Reliability test runs

Foundry Run No.	Metallization Run No.	Topside Material	Package Type
1	C1	Oxide	CPGA
2	P1	Nitride	PPGA
3	C2	Oxide	CPGA
4	C3	Oxide	CPGA
5	P2	Nitride	PPGA
6	C4	Oxide	CPGA

Because of the expense and the time involved in preparing the test structures, burn in boards, and failure analysis equipment, the experiment was designed to test out the gate oxide reliability relationship between Wafer Level Reliability and long term burn in life tests. By concentrating on a single failure mode, tests and test structures can be developed to efficiently monitor for the gate oxide reliability. The same procedure can then be repeated for electromigration and hot carrier reliability monitoring at the wafer level.

7.3.1 Wafer Level Reliability Tests

The wafer reliability tests were run on specific test structures on a test chip adjacent to a product logic chip. The results from the WLR test chip can be used to infer reliability data from the same product die on the same wafers. On the test chip, there are various test capacitors that were used for the Wafer Level Reliability tests. These test capacitors were of the largest area possible to be sensitive to gate oxide failures. The capacitor structure consists of doped polysilicon over gate oxide over either N or P doped silicon wells. The capacitors are large. The four capacitors used in the VRAMP tests are: two capacitors, both P and N type substrates, with an area of 0.001 mm^2 and two capacitors, both P and N type substrates, with an area of 0.005 mm^2. The test capacitors were stressed by using VRAMP tests to determine gate oxide reliability.

The procedure used for the VRAMP tests was the same as the standard JEDEC VRAMP method. In the JEDEC method, the voltage is ramped at a rate of 0.3 MV/cm/sec, to a maximum of 30 MV/cm. The current density flow through the capacitor is monitored and the failure is reached when the current density flow is equal to or greater than 30 A/cm^2. Table 7.2 below shows the results of the VRAMP tests normalized by comparing the test failures to the production standard failure rate. This normalization of test results to standards is done through out this chapter.

Table 7.2 VRAMP normalized data

Run No.	BREAKDOWN VOLTAGE RANGE							
	0 V	0+V to 5V	5V to 10V	10V to 15V	15V to 20V	20V to 25V	25V to 30V	>30V
C1	0.0%	4.7%	8.6%	1.4%	35.5%	50.2%	0.0%	0.0%
C2	0.9%	0.0%	0.0%	0.0%	0.0%	99.1%	0.0%	0.0%
C3	6.5%	0.0%	0.0%	0.0%	18.5%	75.0%	0.0%	0.0%
C4	33.4%	0.4%	0.0%	0.0%	0.7%	14.8%	50.7%	0.0%
P1	38.8%	0.1%	0.0%	0.0%	0.6%	60.4%	0.0%	0.0%
P2	24.7%	0.0%	0.0%	0.0%	1.1%	74.2%	0.0%	0.0%

The VRAMP data can be categorized into three categories: immediate failures, other failures and no failures (passing die). An immediate failure on a VRAMP test occurs when the gate oxide breaks down as soon as any potential is applied. The production die from any lots with immediate failures would probably have either yield or infant mortality failures.

The other failures refer to test die that have a gate oxide that can hold a potential across it, but the gate oxide breaks down at less than 15 volts. This type of failure in the VRAMP data indicates an early lifetime failure in production die. Consequently, the passing die are determined to have good gate oxide when the VRAMP breakdown is greater than 15 volts.

The results of the VRAMP tests, grouped into these three failure categories are shown below in Table 7.3.

Table 7.3 VRAMP normalized failure results

Run No.	Normalized Immediate Failures	Normalized Other Failures	Normalized Total Failures
C1	0.0%	14.7%	14.7%
C2	0.9%	0.0%	0.9%
C3	6.5%	0.0%	6.5%
C4	33.4%	0.4%	33.8%
P1	38.8%	0.0%	38.8%
P2	24.7%	0.0%	24.7%

In addition to the VRAMP monitoring of the large capacitors on the test die, other tests were done to determine gate oxide reliability. On the large test capacitors, a QBD test was done. QBD refers to the charge required to break down a dielectric or Charge to **Break Down.** In this test, a constant current density of 0.121 A/cm2 is applied to the gate oxide and the time to breakdown of the gate oxide is recorded. From the time to breakdown and the current applied, the charge can be calculated. A pass was declared when the gate oxide is capable of holding a charge greater than 5 C/cm^2. This test also showed immediate failures. An immediate failure for a QBD test is the same as that for a VRAMP test, the gate oxide has broken down before any current can be applied to it. The results of the QBD failures are normalized to a standard QBD failure and are presented in Table 7.4 below.

Table 7.4 QBD normalized test results

Run No.	Normalized Immediate Failures	Normalized Other Failures	Normalized Total Failures
C1	22.8%	0.0%	22.8%
C2	15.8%	0.0%	15.8%
C3	5.3%	15.8%	21.1%
C4	0.0%	0.0%	0.0%
P1	0.0%	0.0%	0.0%
P2	5.3%	5.3%	10.6%

Another test for gate oxide reliability was done on the scribeline parametric test transistors. The BVOX test was done on the small gate oxide of the parametric test transistor. BVOX refers to **B**reakdown **V**oltage of **OX**ide. This test is also done at the wafer level, using a test transistor structure in the scribe line. Both the BVOX and the QBD tests were done at the wafer level as a comparison to the VRAMP tests. The BVOX test is similar to a VRAMP test, except that the BVOX test did not follow the JEDEC ramp standard. A pass is determined when the gate oxide breaks down at a voltage greater than 12 volts. Table 7.5 shows the results of the BVOX tests, normalized to a standard BVOX yield.

Table 7.5 BVOX normalized failure results

Run No.	Normalized BVOX N Failures	Normalized BVOX P Failures
C1	2.7%	0.3%
C2	0.0%	0.0%
C3	0.0%	0.0%
C4	0.3%	0.0%
P1	0.0%	0.0%
P2	0.0%	0.0%

The test transistor that the BVOX tests were done on has a gate oxide area of $20\mu m^2$ or $0.000020\ mm^2$. This gate oxide area is much smaller than the gate oxide used for the VRAMP tests on the Wafer Level Reliability test chip.

7.3.2 Wafer Sort and Final Test

The product die from the test runs went through the standard wafer sort operation. The yield normalized to production yield standards are given below in Table 7.6.

Table 7.6 Normalized wafer sort yields

Run No.	Total No. Die	Normalized % Yield
C1	370	151.3%
C2	198	111.1%
C3	198	75.8%
C4	198	0.0%
P1	198	104.8%
P2	198	75.8%

Notice that the run, "C4" did not yield at all at wafer sort. This means that no die were sent on to be assembled and subsequently, final tested and burned in.

The yield results of the final test operation on the packaged production parts is given as a percentage of the production final test yield standard. These yields are given in Tables 7.7 and 7.8. Additional test runs P3, P4 and P5 are included in the final test results for information purposes. These runs went through the metallization processes as the same time as the other test runs shown previously.

Table 7.7 Normalized final test yields

Run No.	Total I.C.'s	Normalized Yield
C1	167	96.3%
C2	104	101.1%
C3	110	103.1%
C4	N/A	N/A
P1	110	105.3%
P2	110	102.5%
P3	57	99.7%
P4	104	99.2%
P5	106	99.3%

Table 7.8 Average normalized final test yields

Number of Runs	Package Type	Total I.C.'s	Average Normalized Yield
3	Ceramic	214	102.1%
5	Plastic	487	101.4%

7.3.3 Reliability Stress Tests

The product die which were packaged and final tested were then stressed in the traditional lifetest manner. By testing on the adjacent product die next to the Wafer Level Reliability test die, information about the relationship of the Wafer Level Reliability test structures to actual product reliability can be determined. Two types of packages were used for the reliability burn in stressing and testing. These were CPGA (Ceramic Pin Grid Array) and PPGA (Plastic Pin Grid Array). The die that were packaged in the CPGA packages had oxide passivation on them and the die that were packaged in the PPGA had nitride passiviation on them. The reason that the CPGA and the PPGA packages were chosen is that there was available burn in boards for these package options for the High Temperature Dynamic Lifetest.

For the Ceramic Packages, (CPGA) the following tests were done: one-third of the total units (parts) were subjected to 1000 hours of 150℃ HTS (High Temperature Storage), one-third of the total units were subjected to 1000 hours of 175℃ HTS, and one-third of the total units were subjected to 100 hours of 125℃ at 6 Volts HTDL (High Temperature Dynamic Lifetest).

The Plastic Packages, (PPGA) had the following tests done: one-quarter of the total units (parts) were subjected to 1000 hours of 150℃ HTS (High Temperature Storage), one-quarter of the total units were subjected to 1000 hours of 125℃ at 6 Volts HTDL (High Temperature Dynamic Lifetest), one-quarter of the total units were subjected to 1000 hours of 85℃ at 6 Volts at 85% relative humidity HTHL (High Temperature/Humidity Lifetest) and one-quarter of the total units were subjected to 96 hours of 121℃ at 15 PSIG PPT (Pressure Pot test). The Pressure Pot stress test is similar to the High Temperature/Humidity Lifetest in that the high pressure is used to force moisture into the plastic packages. If there is a package that is not sealed, this test precipitates failures.

Table 7.9 shows that amount of product that was subjected to these reliability stress tests.

Table 7.9 Reliability stress test configuration

Run No.	Stress Test	No. Hours	No. I.C.'s
C1	HTDL	1000	45
	HTS @ 150 °C	1000	45
	HTS @ 175 °C	1000	45
C2	HTDL	1000	32
	HTS @ 150 °C	1000	32
	HTS @ 175 °C	1000	32
C3	HTDL	1000	32
	HTS @ 150 °C	1000	30
	HTS @ 175 °C	1000	32
P1	HTDL	1000	25
	HTS @ 150 °C	1000	25
	HTHL	1000	25
	PPT	96	25
P2	HTDL	1000	25
	HTS @ 150 °C	1000	25
	HTHL	1000	25
	PPT	96	25

Table 7.10 shows the results of the long term burn in tests. The results are normalized with respect to the expected number of failures normally found in ongoing monitor lots. Therefore, a normalized failure of 100% equals the standard. Any failure over 100% is in excess of the normal standard.

Table 7.10 Reliability stress test results

Run Number	Reliability Stress Test	No. Hours	% Normalized Failures	Failure Analysis Result
C1	HTDL	48	333.5%	Gate Oxide Pinhole
	HTDL	500	119.0%	Gate Oxide Pinhole
	HTS @ 150 °C	500	111.0%	Testing Failure [1]
	HTS @ 150 °C	1000	135.0%	Metal 1 Slit Void
	HTS @ 175 °C	1000	111.0%	Testing Failure [1]
	HTS @ 175 °C	1000	682.2%	Metal 1 Slit Void
C2	HTDL	48	156.3%	Metal 1 Slit Void
	HTDL	500	161.3%	Metal 1 Slit Void
	HTDL	1000	1000.0%	Metal 1 Slit Void
	HTS @ 150 °C	500	312.5%	Open Via
	HTS @ 175 °C	500	312.5%	Open Via
	HTS @ 175 °C	1000	1500.0%	Metal 1 Slit Void
C3	HTDL	168	156.3%	Open Via
	HTS @ 150 °C	500	166.7%	Open Via
P1	HTDL	1000	200.0%	Open Via
P2	HTDL	1000	200.0%	Open Via

NOTE: 1.The reliability failures that were determined to be "Testing Failures" are not real reliability failures. These failures are due to incorrect functional testing during the reliability stressing. During the failure analysis, these parts were retested and determined to be functional ICs.

The above table, Table 7.10 shows the results of the individual burn in tests. The entire lot failures from these tests are summarized by failure category in Table 7.11.

Table 7.11 Reliability stress test results summary

Run Number	% Normalized Failures	Failure Analysis Result
C1	259.5% 148.0% 74.0%	Metal 1 Slit Void Gate Oxide Pinhole Testing Failure - not real
C2	885.4% 208.3%	Metal 1 Slit Void Open Via
C3	106.4%	Open Via
P1	50.0%	Open Via
P2	50.0%	Open Via

7.4 Data Analysis

In looking at the previous tables, the results from the more traditional tests can be compared to the Wafer Level Reliability test results. The results of the wafer level tests on the gate oxides - VRAMP, QBD and BVOX - will be compared with the traditional yield and reliability results - wafer sort yields, final test yields and the reliability lifetime stress tests.

The first Wafer Level Reliability test to be examined is the Break Down Voltage Oxide (BVOX) parametric test. This test was conducted on the scribe line transistors. The results showed no correlation to either the yield or the lifetime reliability stressing. This result is not surprising. The BVOX test is done during the electrical parametrics test on the same scribe line transistors. These test transistors are the same size as the minimum dimension transistors that are used throughout the gate array logic and have a minimal gate oxide area. This small area is insensitive to anything less than a catastrophic level of gate oxide defects. From the results in Table 7.5, the small gate oxide area essentially shows no defects and therefore is insensitive to the level of defects that were shown to exist by the VRAMP tests. This is an example of what parametric testing detects, gross processing problems. BVOX results show only minor problems with two runs - C1 and C4. The yield results show that C1, did not have yield problems yet run C4 did not yield at all. Run C1 did not have yield problems but did have failures in lifetest later determined to be gate oxide pinholes. These gate oxide pinholes were determined by failure analysis to be a result of Electrostatic Discharge (ESD) damage. Because there is not a significant number of failures, it is difficult to separate the failures from the natural variation in the test failures. The gate oxide area is not large enough on the test transistor for the scribe line test transistor to be used as a Wafer Level Reliability monitor structure. The scribe line transistors are used in parametric testing as gross process monitors to detect large errors in processing. Parametric testing will detect such processing errors as a missing implant step.

A similar analysis can be made from the charge to break down (QBD) results. These tests were also done on a few parametric test transistor sites on the wafer. Again, some gross results can be seen from these tests. Only a few tests were done because this type of test is a medium stress test and takes a long time to complete. For good gate oxides, QBD takes a number of minutes for the gate oxide to break down. Therefore, only a small sample of ten sites per run was manually tested.

With a larger gate oxide area, there is a higher probability of detecting defects in the gate oxide. The VRAMP test were done on the large capacitors on the Wafer Level Reliability test chip. With the ramped voltage tests, (VRAMP), there is a relationship between the VRAMP failures and both the wafer sort yield and reliability lifetests. In looking at the yield results, the ceramic runs (C1 through C4) that went through the wafer sort at the same time were compared with the initial failure results of the VRAMP tests. An immediate failure in VRAMP was a result of a failure so severe in the gate oxide that the gate oxide in the test capacitor is incapable of holding any charge at all. This means that time 0, the voltage across the test capacitor to the N or P well is 0 volts. This lack of capacitance is usually indicative of serious processing problems.

Severe processing problems usually cause yield loss. Therefore, in comparing the normalized VRAMP initial failure results with the normalized wafer sort yield results in the graph shown in Figure 7.6.

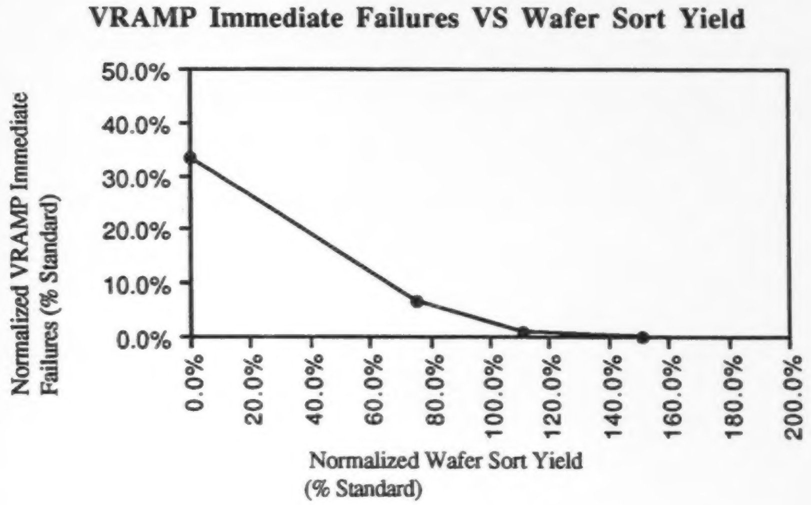


Figure 7.6 Wafer dielectric failure versus wafer sort yields

As the number of sites of test die that fail due to immediate failures increases, the lower the yield of the production die is. This can be compared to a defect density. With more gate oxide failures, the higher the defect density level of this particular failure is. That is, the more gate oxide area is damaged, until no good product die are left in a particular test run. The relationship is a geometric one, dependent upon the area affected. This is a similar result as the various yield model would predict. When the amount of gate oxide failures as shown by the VRAMP test reaches more than 33% normalized immediate failures as compared to the standard failure rate, the adjacent product die on the same test wafers have no good functioning die at wafer sort.

In analyzing the VRAMP test data, not all the failures that occurred were of the immediate failure type. The other failures occurred when the gate oxide broke down before 15 volts was reached during the VRAMP test. These VRAMP failure results are indicative of a weak gate oxide, not a failed gate oxide. The gate oxide breaking down at less than the normal voltage is more an indication of a reliability problem than a wafer sort yield problem. A weak gate oxide will still allow the die to work initially but a weak or contaminated gate oxide will cause an early lifetime wearout.

Therefore, in looking at the lifetime reliability results, the failures that were determined to not be related to the gate oxides will not be considered in a comparison with the VRAMP non-immediate failures. From Table 7.10, the only run that had gate oxide reliability lifetest failures was run C1. The other runs are considered to have no reliability lifetest gate oxide failures. The graph of the normalized VRAMP failures compared to the normalized reliability lifetest failures is shown in Figure 7.7.

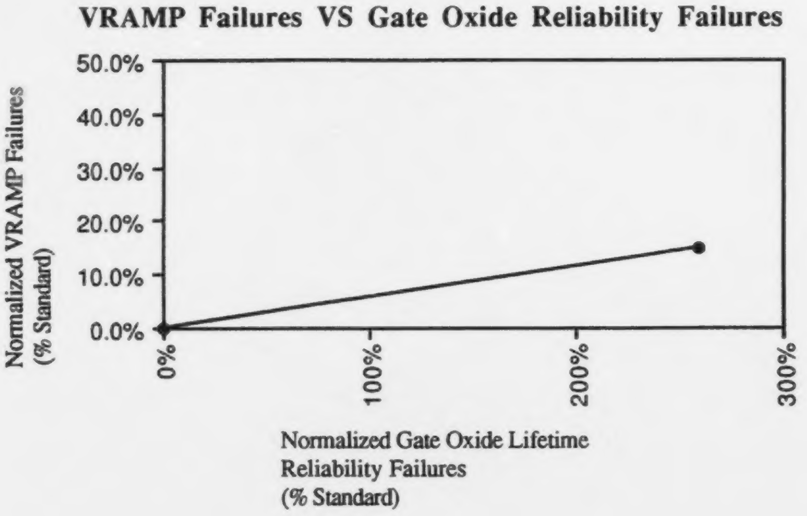


Figure 7.7 Wafer failures versus gate oxide lifetime reliability failures

The only run that showed weak gate oxide results from the VRAMP tests is run C1. Run C1 was also the only run that had failures due to gate oxide pinholes. The other runs (C2 and C3) had no gate oxide failures and essentially no VRAMP failures. Runs C1 and C2 have their data point overlap at the origin of the above graph.

The gate oxide pinhole failures were determined to be due to ESD damage. This was determined through failure analysis of the parts that failed the reliability lifetime burn in tests. The Wafer Level Reliability JRAMP stress test was able to predict a reliability problem with this particular lot. The traditional reliability stress tests clearly show the gate oxide problem.

The results from final test do not show any specific failure modes or correlation to any wafer reliability tests, except for a lower final test yield for run C1. This is interesting to note because the final test is another functional test, similar to the wafer sort operation. If the wafer sort operation is done correctly, there is a small amount of failure at final test due to either assembly induced defects or die that have infant mortality problems. Run C1 showing a lower final test yield along with having a weak gate oxide is not unexpected. Outside of the usual assembly related defects, there would be some die in run C1 that barely passed the wafer sort testing operation due to the weak gate oxide. These additional die at final test failed due to the stressing of the heat cycles encountered in the assembly processes and the electrical stress of final functional testing. The normalized VRAMP test results compared to the normalized final test yields is presented in the graph in Figure 7.8, below.

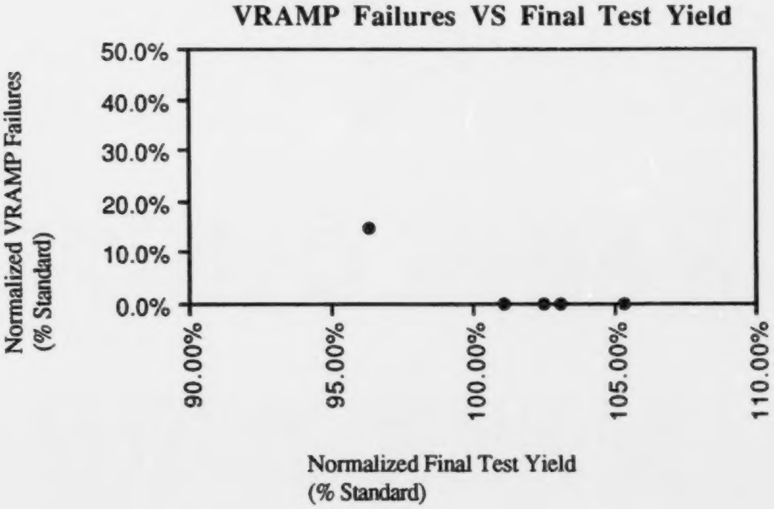


Figure 7.8 Wafer failures versus final test yield

From the above graph, there appears to be a decrease in the final test yield with an increase in the failure of the VRAMP tests. The final test yield has natural variance associated with it. If the wafer sort test has been done correctly, there will be a small percentage of failures associated with "escapes" from wafer sort in the final test yield. The final test yield will mostly reflect failures due to the assembly processes. The above graph shows the final test yield of five of the six lots - C1, C2, C3, P1 and P2. With no yield on run C4, there were no good die to assemble and consequently, no parts (ICs) for final test.

The only run that had a significant failure during the Wafer Level Reliability VRAMP test also had the lowest final test yield, run C1. This final test yield was outside of the normal variance seen in production at final test. Although no failure analysis was done on the failed run C1 final test yield loss, the yield was significantly less than the normal standard assembly ceramic yield. The standard yield is shown as a normalized final test yield of 100%. It can not be determined that the yield loss was due to the weak gate oxide without the failure analysis, but this would seem to be a reasonable assumption due to the Wafer Level Reliability data.

7.5 Wafer Level Reliability

The question still arises with the results presented on run C4, why was it not detected until the wafer sort operation? The answer to this question is that the manufacturing processes entered an out of control condition that was not detected. This is why the gate oxide reliability is only one factor that must be considered when determining what failure modes to monitor for with Wafer Level Reliability.

To develop a set of WLR structures to be used in monitoring the reliability of a manufacturing process, the major failure mechanisms must first be known and understood. By identifying the major failure mechanisms, test structures can then be designed and developed to test for these failures. The purpose of using Wafer Level Reliability is to have the failures known as early in the process as possible. To be proactive, rather than reactive.

Gate oxide wafer level test structures are just one type of test structures to test one major failure mechanism for reliability failures. In examining the results of the burn in lifetime tests in Table 7.11, there were also failures for open vias and metal 1 slit voids. These are failures that possibly could be detected early through the use of Wafer Level Reliability monitoring.

Wafer Level Reliability is not just a single type of test structure or failure mode. The slit void failure is a first layer metal stress void failure. This can be detected by use of electromigration test structures such as the SWEAT test structure. The monitoring of via chain electromigration structures may have detected the via failures. Further work is needed in determining exactly which structures will accurately detect which failure modes.

Certain failures may not be detected by test structures that were designed to detect problems. For example, a particular via problem may be a result of sensitivities of a particular design to a particular process. The test structure in question may not be designed to have a structure that is sensitive to a particular failure. Even though there are test structures and procedures in place to detect via problems, the problems that are structurally sensitive can go undetected by some Wafer Level Reliability test structures.

To over come this, the development of Wafer Level Reliability test structures must be a constant ongoing process. As new technologies are developed with smaller and different designs with different materials, new failures will result. As processes are changed, different process sensitivities will result causing new reliability failure modes. Not all of the failures that can occur on a particular design can be detected. Wafer Level Reliability is not intended to replace the traditional monitoring of lifetime through the various accelerated burn in tests. However, the advantage of Wafer Level Reliability is that it requires much less time to do than the traditional lifetime burn in tests. Tests are designed to take less than a minute for Wafer Level Reliability tests as compared to the traditional lifetime tests. Another advantage of Wafer Level Reliability tests is that it does not destroy product. Lifetime burn in stressing destroys the product it is testing. Wafer Level Reliability test structures can be put into the scribe line area next to every product die. The importance of this is that information on the reliability of product can now be gathered without destroying the product itself. For specific reliability concerns, product can now be monitored and data can be gathered about the reliability of the product.

Because the Wafer Level Reliability tests can be done in a short amount of time, it is now possible to do reliability monitoring on every single wafer that is produced. This amount of sampling will increase the confidence level of the reliability of the final product. The outgoing quality of the product can be statistically guaranteed to be at a certain confidence level for a given number of different reliability failure modes.

A question still arises about how one obtains lifetime data from Wafer Level Reliability tests and test structures. While it is not feasible to gather complete lifetime data of a part that have many different reliability failure modes from a number of different test structures that are designed to only monitor single failure modes, it is feasible to gather indications of the product lifetime from these test structures. The test structures will only show reliability failures that they are designed to detect. If there is a reliability failure that no test structure exists to detect, it will go undetected. The procedure is to have Wafer Level Reliability a vital part of an overall quality program.

7.5.1 Statistical Process Control

Constant sampling of lots through traditional lifetime burn in tests will help to identify the major reliability failure modes. From these failure modes, test structures can be used or designed to detect the failures. The sensitivity of the test structures and the testing methodology can be determined by experiments such as the gate oxide Wafer Level Reliability testing done here. Statistical Process Control (SPC) principles are important in determining the relationship of Wafer Level Reliability results to burn in lifetime reliability results. When the manufacturing process is in control, there will be some small variance in both the Wafer Level Reliability tests and the lifetime burn in tests. SPC can be used to show the difference between the natural variance of a Wafer Level Reliability test and an out of control condition that requires attention. The implementation of Statistical Process Control in Wafer Level Reliability monitoring is outlined below.

Most of the Lifetime stress tests usually involves at least 1000 hours to complete the test. This means that very little of the actual production lots can be sampled for reliability testing. This is also very expensive to do and destroys product ICs in the process. To obtain enough samples to guarantee product lifetime is also usually prohibitively expensive. The answer to this problem is to use accelerated Wafer Level Reliability testing. The key to using Wafer Level Reliability methodologies as lifetime monitoring is to correlate these tests to the results of burn in analysis and use SPC to determine out of control conditions.

The correct way to determine the Median Time To Failure (MTTF) and to have enough sampling of product to insure product reliability is to develop an ongoing comparison of Wafer Level Reliability monitoring with lifetime stress testing. Once a correlation between the MTTF of the burn in tests and the accelerated test is done, the relationship between the Wafer Level

Reliability test methodologies and the lifetimes of the particular reliability is known. Once this work is done on a mature process, a base line of results will be established. This "base line" is a known set of correlations between the Wafer Level Reliability testing results and the Burn in MTTFs. For example, if a 10^{10} hour result of a product had a SWEAT test structure last for 30 seconds before failure, then the 30 second SWEAT result could be considered for a production process as a base line monitor. Through enough test and statistical analysis, a base line for other reliability failure modes can be established. For an in control, mature process, these base line results are the results that are expected from normal Wafer Level Reliability testing.

This base lining of the process leads to the next part of an overall Wafer Level Reliability program. Once the base line is established, it can be used to monitor the day to day production of a VLSI manufacturing facility. By using the principals of Statistical Process Control, SPC, a process monitor (PM) can be set up. The PM would involve various test structures and test methodologies such as SWEAT or BEM, VRAMP or JRAMP, and a hot carrier test methodology. As long as the results of the tests remain within control, that is, within the normal distribution of results, the product will have an acceptable lifetime for the particular reliability failure mode.

If the tests deteriorate below the normal distribution limits, this indicates an out of control condition and product that would have reliability problems. Action can then be taken to identify what has changed in the manufacturing process, and the corresponding problem can be rectified before any more product is affected.

To look at an out of control condition, the example of an electromigration SWEAT test production monitor will be used. This SWEAT monitor can be any monitor of a metal layer electromigration. For the purposes of illustration, a Metal 1 monitor will be used.

The graph below shows the mean time to failure of a typical Metal 1 production monitor SWEAT test on product wafers done on a lot by lot basis:

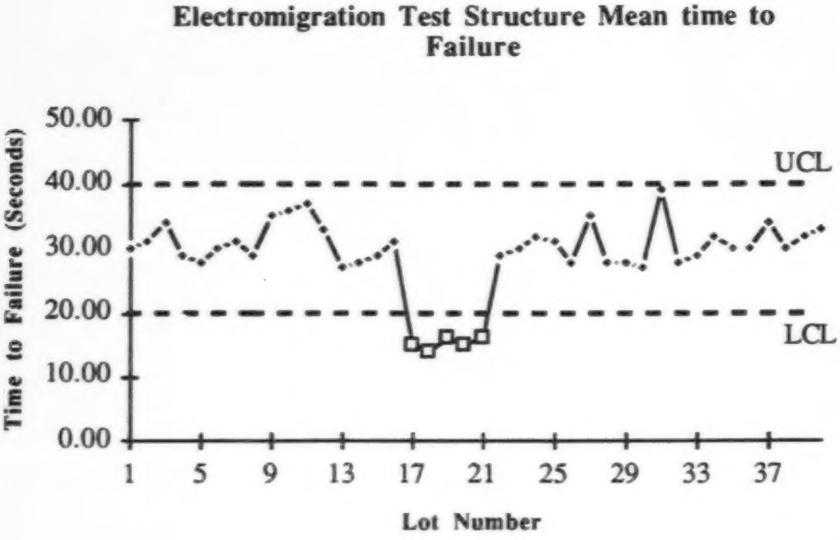


Figure 7.9 SWEAT electromigration control chart

The example electromigration Production Monitor (PM)'s would be tested and a lot average is plotted on a chart like the one shown above. The mean time to failure remains statistically in control between the UCL (Upper Control Limit) and the LCL (Lower Control Limit). When lots fall below the LCL, action must be taken to find out where the manufacturing process is out of control before more defective product is manufactured. Even through the lots that are below the LCL do not have their actual electromigration MTTF known, the results from the correlation experiments clearly show that there is a problem with their lifetime due to electromigration. This is an example of part of an entire reliability and quality manufacturing monitoring program. Similar Wafer Level Reliability monitors would be installed for all the major reliability failure modes.

By using this principle of Statistical Process Control, the reliability of a given process can be monitored. The various wafer reliability results can then be related to a given lifetime failure. When a process does go out of control, the Wafer Level Reliability test will show this. The particular test will fail. The lifetime affect will not be known, but it will be known that the lifetime is being affected and steps can be taken to bring the process back into control. This is perhaps the greatest advantage to Wafer Level Reliability. It is proactive, rather than reactive. Action can be taken when the process goes out of control and affected product can be scrapped out of production. With the traditional lifetest, problems could go undetected for months.

This is because of the length of the traditional lifetime burn in tests take. Not only does it take 1000 hours or more to stress the test parts, there is additional time of testing and retesting the parts during the lifetest plus the additional time to do the failure analysis that determines the exact failure mode. Even after all of this testing, some failures can be missed. Because the traditional lifetime tests consume a large amount of resources, the ongoing Lifetime burn in monitor sample volumes are usually small when compared with production volumes. It is therefore possible for an intermittent reliability problem to go undetected by using only the traditional Lifetime burn in tests for reliability screening.

The sooner a reliability problem is detected, the less the effect upon the product and customers is felt. By testing for reliability within the manufacturing line by the use of Wafer Level Reliability rather than after, reliability problems can be resolved before a large amount of product is affected. The flow chart in Figure 7.10 gives an example of where Wafer Level Reliability tests could be incorporated into a semiconductor manufacturing process flow.

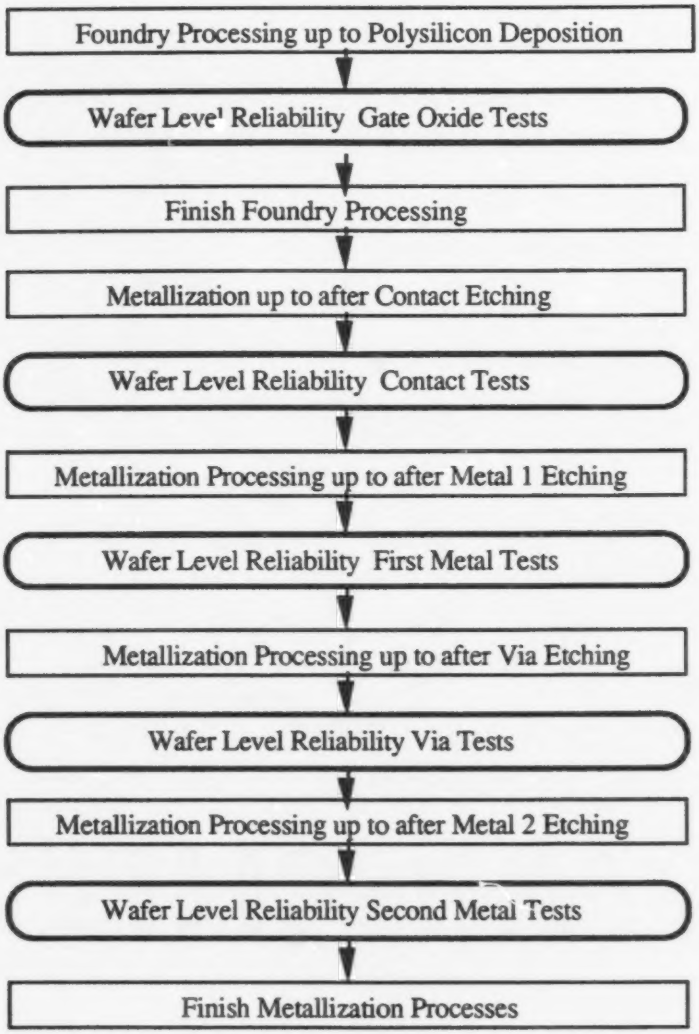

Foundry Processing up to Polysilicon Deposition
Wafer Level Reliability Gate Oxide Tests
Finish Foundry Processing
Metallization up to after Contact Etching
Wafer Level Reliability Contact Tests
Metallization Processing up to after Metal 1 Etching
Wafer Level Reliability First Metal Tests
Metallization Processing up to after Via Etching
Wafer Level Reliability Via Tests
Metallization Processing up to after Metal 2 Etching
Wafer Level Reliability Second Metal Tests
Finish Metallization Processes

Figure 7.10 Wafer Level Reliability Process Monitor flow chart

As a wafer is processed, time and resources have been spent on the manufacture of the wafer. The earlier problems can be detected, the earlier the process can be corrected and the less product is affected. Reliability monitoring within the manufacturing processes is now possible through the use of Statistical Process Control, Wafer Level Reliability development and Lifetime burn in data correlation.

7.5.2 Product Lifetime Modeling

One of the major goals of a semiconductor reliability program is to gain information about the lifetime of Integrated Circuits. Most semiconductor lifetimes provided by manufacturers are derived from the results of long term burn in lifetime testing. Wafer Level Reliability can not by itself predict product lifetime.

WLR testing can be used in a process monitoring function as described previously, by using the principles of SPC monitoring to detect when the reliability of a product is outside of the normal variance seen in the WLR test data. If the individual results of these tests are applied to the Arrhenius equation, one will derive a number for the lifetime of a major reliability failure mode. This lifetime number would not be accurate for the following reasons. Wafer Level Reliability is done on a test structure that is specifically designed to cause failures for that one mode only. The test structure is different that the structure of the actual production circuitry. The lifetime data obtained is a representation of the lifetime of the test structure circuit under normal operation, not the associated product IC.

The other difficulty is that the WLR tests are highly accelerated. Even a small error in measurement or drift in equipment calibration can cause a large error in lifetime calculations. It is almost impossible to determine lifetimes of 10 years or so from a test that takes only 30 seconds.

Yet the desire is to take the data that Wafer Level Reliability provides and apply it to product lifetimes. WLR testing provides a possibility of collecting a large amount of data quickly. To use this data in the lifetime calculations of product predictions would enable the lifetime effects of process problems to be determined. With a limited amount of sampling that takes place with the lifetime burn in tests, information about the lifetime effects of process problems is difficult to obtain. Small sample sizes of reliability monitoring lots for lifetime burn in can miss process problems. With WLR structures present in scribe line structures adjacent to production die, data can be obtained from every wafer that has a potential process problem.

The issue is still relating the data from WLR test data to product lifetime. The methodology illustrated in this chapter to determine the relationship between the VRAMP results and the gate oxide reliability is an example of the methodology of relating the WLR test data to the lifetime burn in results. Through additional correlation of other WLR test structures such as SWEAT or hot carrier structures to lifetime burn in results, empirical relationships can be established between Wafer Reliability tests and actual product lifetimes. This methodology is how WLR test can be used to determine product lifetimes. The lifetime of product can be modelled by first understanding the empirical relationship between product lifetime burn in tests and Wafer Level Reliability data.

WLR test structures can be tested to determine if they actually detect the major reliability failure modes that they are designed to. The test structures and the associated test methodology can then be refined such that the major failure modes are detected. A test chip with these major test structures is then placed next to a production die as illustrated in Figure 7.4. Additional data can then be gathered to determine the relationship between the WLR test results and the lifetime burn in results. The product die adjacent to the WLR test die is subjected to the lifetime burn in test to determine the FIT rates of the major reliability failure modes. Over a period of time, empirical relationships between the major reliability failure modes contribution to the overall FIT rate and the WLR test time can be established. For example, if the normal WLR metal 1 SWEAT test is shown to last for 30 seconds, and this corresponds to a contribution to the overall FIT rate of 2.5 fits, then an empirical data relationship is known. If a particular lot experiences a major process problem such as corrosion, the metal 1 SWEAT survival time may drop to 20 seconds. The associated product would not be used in its intended application, but would be used to determine the effect on product lifetime of the metal 1 corrosion problem. This would then establish another data point in the WLR and lifetime relationship. Over time enough empirical data has been gathered, a model of the major reliability failure modes contribution to the overall FIT rate can be established. An example of what this limited reliability model is given below in Table 7.12.

Table 7.11 Reliability stress test results summary

Major Failure Modes	**FIT Rate Contribution**
Gate Oxide	1.7
Hot Electron	0.8
Contact Electromigration	0.3
Metal 1 Electromigration	0.7
Metal 1 Stress Voids	0.2
Via Electromigration	0.3
Metal 2 Electromigration	2.3
Metal 2 Stress Voids	0.9
Passivation Pinholes	0.0
Assembly Static Discharge	0.4
Not Determined	1.3
TOTAL FAILURE RATE	**8.9**

The not determined data would be due to failures that could not be identified or not monitored for in a WLR test structure. This model would be updated as the process changes and evolves. In addition to new Wafer Level Reliability tests and test procedures, new lifetime stress tests that precipitate the lifetime failures would also be developed. From this model, quick information about the lifetime can be obtained quickly though the use of WLR data. The traditional lifetime burn in tests would be used mostly for the evaluation and updating the WLR reliability model and less for product lifetime evaluation.

7.6 Summary

The relationship between the Wafer Level Reliability test data and the lifetime burn in results has been demonstrated through the testing of ASIC CMOS gate oxide. VRAMP WLR testing was shown to predict lifetime and yield failures on product die adjacent to the test chips. Through the use of High Temperature Storage and High Temperature Dynamic Lifetime stressing, failures were found on a lot that had VRAMP failures. With severe WLR VAMP immediate failures, the wafer sort yield was also seen to be affected.

The procedure outlined in this gate oxide test shows how Wafer Level Reliability can be used at various process steps as a production monitor for reliability. This WLR production monitoring within the various process steps, this in process monitoring, can identify the major reliability defects within the process, not months after the problem. Data can be gathered rapidly and changes can be made through the principles of Statistical Process Control, SPC.

Further research is required to develop the Wafer Level Reliability test result relationship to lifetime burn in data results. By following the same procedure demonstrated on the gate oxide reliability for other major reliability failure modes, empirical data can be gathered to develop a Wafer Level Reliability lifetime model. This model can then be used to estimate the product lifetime much quicker than the existing lifetime burn in procedures. Lifetime information about the impact of process changes can be assessed quickly, rather than in a period of months as it takes today.

The only way to insure that the model remains accurate and can roughly estimate the lifetime of product, is to have an ongoing Wafer Level Reliability monitor program. The purpose of developing this program is to continually update the model, as well as identify new failure modes. With the identification of new failure modes, new WLR structures and test procedures can be developed and old ones updated to detect the new failure modes. The Wafer Level Reliability data can be empirically related to adjacent burn in lifetime failure rates. The model can then be updated to reflect new data and improve the accuracy of the model.

Further research and development is required in the area of Wafer Level Reliability to develop a model that can give a good indication of product lifetimes. Even with a good model, there will be limitations to the prediction accuracy. The model will only be able to show failures in the major failure modes, not all of the possible failures. With more development of this Wafer Level Reliability model, it may be possible to estimate product lifetime data from in process monitors. This would mean specific wafers could have their reliability assessed by testing the WLR structures on them.

CHAPTER VIII
CONCLUSIONS

The semiconductor reliability research activities conducted in this thesis were directed at providing a "state of the art" knowledge base and an evaluation of the new and emerging field of Wafer Level Reliability, specifically for Application Specific Integrated Circuits (ASIC). Semiconductor designs and processes are continually undergoing rapid changes in complexity, performance and size in response to increased competition and the economic pressures of the industry. Semiconductor manufacturers are challenged to produce integrated circuits that are inexpensive yet more reliable than ever before. New semiconductor processes, testing methodologies and procedures are being developed to accommodate these changes by increasing the amount of reliability assurance testing through Wafer Level Reliability Testing, the subject of this thesis. Presently, there is very little, if any, published material on this new area.

Initially, the fundamentals of semiconductor manufacturing processes (e.g., photolithography, ion implantation and diffusion, thin film deposition, etching, etc.) were discussed and illustrated is some detail to form a part of the knowledge base required to understand the possible failure mechanisms and defects that can occur due to process irregularities. These fundamentals were required to demonstrate the dominant failure mechanisms that affect the reliability of semiconductor devices and to reveal the importance of Wafer Level Reliability for the evolving semiconductor industry. Process anomalies were shown to significantly affect the performance and lifetime of integrated circuits.

Yield and reliability are key performance indices of semiconductor manufacturing processes that define the quality of integrated circuits during the various stages of their fabrication. Detailed discussions of the various yield concepts (e.g., line yield, wafer sort yield, assembly yield, final test yield) were presented to reveal how today's semiconductor manufacturers monitor the quality of their products during the various stages of their manufacturing processes. The basic reliability equations (e.g., reliability distributions, burn in reliability, Arrhenius equation, Eyring's equation, etc.) and the concepts of accelerated lifetime testing were developed and provided as basis for assessing why various test procedures are adopted by the semiconductor industry. The various acceleration factors, not easily found in the literature, were presented and provided a means of estimating the failure rate of integrated circuits operating in different environments. The yield and reliability knowledge base were required for the interpretation of the empirical Wafer Level Reliability test results presented in this thesis.

The detailed fundamentals of electromigration, the most critical failure mechanism as seen by the early life time failures of today's complex semiconductor circuits were presented. The basic equations for defining the MTTF of semiconductor circuits and the critical variables that determine reliability test procedures and screening methodologies were developed and discussed in detail. New test methodologies (e.g., SWEAT and BEM tests) to detect electromigration failures were presented and illustrated.

Wafer Level Reliability (i.e., WLR) Testing fundamentals involving the fabrication and integration of test structures adjacent to production parts on a single wafer and the specific tests for these test structures were discussed and illustrated by unique WLR test structures for ASIC gate array manufacturing. A test chip layout containing the various test structures (e.g., hot carrier, interconnect chains, SWEAT, BPSG oxide dielectric, interdielectric oxide, gate oxide, Metal 1 and 2 SWEAT) was presented. Detailed discussions on the test structures which were designed to detect the three main failure mechanisms in today's semiconductor industry, namely; electromigration, dielectric failures and hot carrier failures was presented. These Wafer Level Reliability test structures applied to an ASIC manufacturing environment have not been published to date.

Empirical test results of traditional life time reliability testing for designed and fabricated wafers containing test circuit die adjacent to ASIC die were presented and analyzed in detail. Wafer Level Reliability test results, VRAMP, QBD, BVOX, wafer sort, final test results and reliability stress experimental test results were presented and analyzed in some detail. It was shown that WLR testing did reveal the major failure mechanisms, many of which were not detected by some of the traditional screening tests. The test results presented in this thesis provide insight into the various reliability tests that are conducted at the end of various process stages in the manufacturing of semiconductor devices.

The effectiveness in detecting known failure mechanisms by Wafer Level Reliability test structures being placed on production wafers was clearly demonstrated in this thesis by an analysis of empirical test data of production runs containing WLR test structures. Wafer Level Reliability provides an economical and time efficient means of detecting major process anomalies and provides a means of rapid feedback to control manufacturing processes that are out of control. Wafer Level Reliability achieves early detection of process anomalies due to the increased number of samples taken per wafer utilized by this testing methodology. Traditional lifetime testing methodologies can take several months to identify major process failure mechanisms while Wafer Level Reliability testing requires significantly less time to identify the same major failure mechanisms. Analysis of the test results also revealed that some of the traditional screening tests failed to detect certain failure mechanisms that were detected by WLR testing. Through the use of in process production monitors, failures can be detected early in the process.

Another major economical advantage of Wafer Reliability testing is that it is a non-destructive test and does not consume product in testing, a significant economical advantage over other testing methodologies. It can be easily incorporated into existing semiconductor manufacturing processes.

A primary question posed by this thesis was: could Wafer Level Reliability testing predict the lifetime of semiconductor devices? WLR testing provides an estimate of product lifetime based on the identification of the major semiconductor failure mechanisms rapidly. It can not replace traditional lifetime testing methodologies which require months of testing to provide more accurate estimates of product lifetime. Future research is required to develop accurate models of the relationships of WLR test results to traditional lifetime testing, before product lifetimes can be estimated by Wafer Level Reliability test results. An example of this modeling was discussed in detail in this thesis.

It is important to remember that no testing methodology can economically identify all the possible semiconductor failure mechanisms. The failure rate of a semiconductor device is dependent upon its operating environment and the stresses placed on the individual ICs. For example, a silicon rectifier operating in different environments has the following failure rates:

Table 8.1 Semiconductor failure rates in different environments

• air conditioned digital computer	100	FITS
• data handling system	100	FITS
• control room	250	FITS
• normal industrial	500	FITS
• shipboard, chemical environment	1000	FITS
• mobile, road, rail	2500	FITS
• portable and bench applications	2500	FITS
• airborne	5000	FITS
• rocket launch	8000	FITS

Semiconductor manufacturers can only provide base failure rates for their products and it is the duty of their customers to determine the failure rate of their products depending upon their utilization.

Wafer Level Reliability is a new and emerging field in semiconductor quality assurance of ASIC technologies. Considerable costs and research efforts have advanced the Wafer Level Reliability field to this stage in its evolution. Considerably more research is still required to develop new test structures that can detect new failure mechanisms in new and existing semiconductor circuits. Models that relate Wafer Level Reliability to traditional lifetime tests need to be developed to increase the accuracy of lifetime predictions through WLR testing. Because today's semiconductor integrated circuits are so reliable (e.g. 1 failure per 10^9 operating hours, i.e. 1 FIT), reliability testing methodologies are changing in order to precipitate failures within the shortest time period possible in order to estimate the reliability of new semiconductor devices.

REFERENCES

1. Tim Turner, "Wafer Level Testing", United Technologies-MOSTEK, Carrollton, Texas.
2. S.M. Sze, W.J. Bertram, et. al., *VLSI Technology*, Chapter 14: "Yield and Reliability", pg. 642, McGraw Hill Book Co., 1983.
3. G.E. Dewitt, *Modern MOS Technology: Process Devices and Design*, Chapt. 14: "Yield and Reliability", pp. 339-355, McGraw Hill Book Co., 1984.
4. H. S. Blanks,"The Temperature Dependence of component Failure Rate", Microelectronic Reliability, Vol. 20, pp. 297-300, 1980.
5. Bryan J. Root and Tim Turner, "Electromigration tests for Production Monitoring",IEEE 23rd Annual Proceedings on Reliability Physics, pp. 100-107, 1985.
6. Various, "Logic Design Manual for ASICs", LSI Logic Corporation, 1989.
7. S.M. Sze, T.E. Seidel, et. al., *VLSI Technology*, Chapter 6: "Ion Implantation", pg. 221, McGraw Hill Book Co., 1983.
8. S.M. Sze, W.J. Bertram, et. al., *VLSI Technology*, Chapter 14: "Yield and Reliability", pg. 603, McGraw Hill Book Co., 1983.
9. Various, "LSI Logic Reliability Manual and Data Summary", LSI Logic Corporation, 1990.
10. Tim Turner et. al., "Wafer Reliability Testing", Reedholm Instruments, Fremont, California, 1990.
11. J.M. Poate, Thin Films - Interdiffusion and Reactions, John Wiley & Sons Inc., 1978.
12. James R. Black, "Physics of Electromigration", IEEE 12th Annual Proceedings on Reliability Physics, 1974, P 142-149.
13. Bryan J. Root and Tim Turner, "Electromigration tests for Production Monitoring",IEEE 23rd Annual Proceedings on Reliability Physics, pp. 100-107, 1985.
14. D. Cook & C. Hong, "Breakdown Energy of Metal- BEM",IEEE 23rd Annual Proceedings on Reliability Physics, pp. 108-114, 1985.
15. A.S. Oates, "Step Spacing Effects on Electromigration" IEEE 28th Annual Proceedings on Reliability Physics, pp. 20-24, 1990.
16. S.M. Sze, L.C. Parrillo, et. al., *VLSI Technology*, Chapter 11:"VLSI Process Integration", pg. 496, McGraw Hill Book Co., 1983.

17. K.T. Chan, J. Hui, P.V. Voorde & H.S. Fu, "Self-Limiting Behavior of Hot Carrier Degradation and its Implication on the Validity of Lifetime Extraction by Accelerated Stress " IEEE 28th Annual Proceedings on Reliability Physics, pp. 20-24, 1990.

18. Thomas Kopley, "Hot Carrier Stress Test Structures for Wafer Level Testing" 1991 International Wafer Level Reliability Workshop, Technology Associates, pp.41-54, 1992.

19. Various, *INTEL Components Quality/Reliability Handbook*, Chapter 4: "Reliability", pp. 4-1 to 4-45, 1986.

20. D.S. Peck and O.D. Trapp, *Accelerated Testing Handbook,* Fifth Edition, Technology Associates & D.S. Peck Consulting Corporation, 1987.

21. Finn Jensen and NielsErik Petersen, *Burn-In* , Chapter 7:"Acceleration of Burn-in Tests, pp. 75-87,John Wiley & Sons, 1982.

22. S.M. Sze, D.B. Fraser, et. al., *VLSI Technology*, Chapter 9: "Metallization", pg. 370, McGraw Hill Book Co., 1983.